DOCUMENT RESUME

ED 323 197 SP 032 584

AUTHOR	Estes, Gary D.; And Others
TITLE	Assessment Component of the California New Teacher Project: First Year Report.
INSTITUTION	Far West Lab. for Educational Research and Development, San Francisco, Calif.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	Mar 90
NOTE	211p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	*Beginning Teachers; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; *Measurement Techniques; Teacher Characteristics; *Teacher Evaluation

ABSTRACT

This assessment component of the California New Teacher Project consists of the development and pilot testing of innovative forms of new teacher assessment. The evaluation of diverse approaches to teacher assessments is intended to identify the most promising ways in which a comprehensive assessment of teacher candidates could inform the certification process and contribute to the quality of teaching. The introduction to this document presents a review of literature on new teachers, focusing on the incompleteness of preservice training, problems of new teachers, and differences between novice and expert teachers. Specific contributions of the spring 1989 round of pilot testing are discussed. The purpose of the pilot testing was to examine in California the functioning of several assessment instruments which are considered to be promising exemplars of innovative assessment approaches. The evaluation of the various components of these instruments (e.g., logistical requirements, prompt materials, scoring criteria, training exercises for assessors) was intended to provide information concerning the strengths and limitations of the assessment approaches which specific instruments represented. This final report and analysis of the pilot test administration describes each of the instruments, and the ease of administration, scoring, content and format, costs, and technical qualities. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

SP

ED323197

ASSESSMENT COMPONENT OF THE CALIFORNIA NEW TEACHER PROJECT: FIRST YEAR REPORT

MARCH 1990

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T. L. Ross
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SP 032 584

FAR WEST LABORATORY
FOR EDUCATIONAL RESEARCH AND DEVELOPMENT
1855 FOLSOM STREET · SAN FRANCISCO, CALIFORNIA 94103

BEST COPY AVAILABLE

ASSESSMENT COMPONENT OF THE CALIFORNIA NEW TEACHER PROJECT:

FIRST YEAR REPORT

Far West Laboratory for Educational
Research and Development

Gary D. Estes
Kendyll Stansbury
Claudia Long

With the assistance from RMC Research Corporation

Patricia Wheeler
Edys Quellmalz

March 1990

ACKNOWLEDGEMENTS

Both the pilot testing and report writing phases of the project benefited from the assistance of many people. Staff from the Commission on Teacher Credentialing and the State Department of Education provided timely support and critical information that enabled the pilot testing to occur during a limited time period. We also benefited greatly from their review of draft reports. The Project Directors of the support projects assisted in explaining the importance of the evaluation of the assessments to potential participants, recruiting teachers, identifying facilities, and suggesting districts to identify additional teachers. Their cooperation and patience made our job much easier. We also wish to acknowledge the contributions of the teachers who participated in the assessments, particularly the teacher who traveled over 100 miles on her birthday to take a two-hour examination.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS i

CHAPTER 1: INTRODUCTION 1.1

Review of Literature on New Teachers 1.2

Incompleteness of Preservice Training 1.2
Problems of New Teachers 1.3
Differences between Novice Teachers and Expert Teachers 1.4

California New Teacher Project 1.5

Assessment Component of the California New Teacher Project 1.6
Rationale and Design of the Assessment Component 1.7

Pilot Testing in 1989 .... 1.9
Terminology Used in the Report 1.10

CHAPTER 2: PILOT TEST DESIGN AND ANALYSIS 2.1

Design of Pilot Tests 2.1

Sources of Instrumentation 2.1
Sampling Plan 2.2

Analysis of Pilot Tests 2.3

Data Collection 2.3
Data Reduction 2.4

Overview of Analytic Categories 2.5

Administration of Assessment 2.5
Assessment Content 2.5
Assessment Format 2.6
Cost Analysis 2.6
Technical Quality 2.6

CHAPTER 3: CONNECTICUT COMPETENCY INSTRUMENT (CCI) 3.1

Administration of Assessment 3.4

Overview 3.4
Logistics 3.5
Security 3.5
Assessors and Their Training 3.8
Scoring 3.10
Teacher, Assessor, FWL, and RMC Staff Perceptions of Administration 3.11

Assessment Content....3.12

Congruence with California Curriculum Guides and Frameworks....3.13
Extent of Coverage of California Standards for Beginning Teachers....3.15
Job-relatedness....3.19
Appropriateness for Beginning Teachers....3.19
Appropriateness across Contexts....3.20

Grade level and subject matter....3.20
Diverse students....3.24

Fairness across Groups of Teachers....3.24
Areas of Most/Least Emphasis....3.25

Assessment Format....3.25

Clarity of the Assessment....3.26
Clarity of Assessment Materials....3.26
Observation Feedback....3.28

Cost Analysis....3.29

Assessor Time....3.29
Training Costs....3.29
Other Costs....3.29
Summary....3.30

Technical Quality....3.30

Development....3.30
Reliability....3.31
Validity....3.32

Conclusions and Recommendations....3.33

Administration of Assessment....3.33
Assessment Content....3.34
Assessment Format....3.35
Summary....3.36

CHAPTER 4: SEMI-STRUCTURED INTERVIEW: SECONDARY MATHEMATICS....4.1

Administration of Assessments....4.3

Overview....4.3
Logistics....4.3
Security....4.5
Assessors and Their Training....4.6
Teacher and Assessor Impressions of Administration....4.7

Scoring 4.7

Math Scoring Indicators and Indicator Elements 4.8
Scoring Process 4.10
Discussion of Scoring System...4.11
Scorers and Their Training...4.12

Assessment Content 4.13

Congruence with California Curriculum Guides and Frameworks.......4.13
Extent of Coverage of California Standards for Beginning Teachers....4.15
Job-relatedness 4.19
Appropriateness for Beginning Teachers....................................4.21

Perceptions..4.21
Performance on assessment...4.22

Appropriateness across Contexts..4.22

Grade level and subject area...4.22
Diverse students...4.22

Fairness across Groups of Teachers 4.23
Appropriateness as a Method of Assessment 4.23

Assessment Format 4.23

Clarity of Assessment...4.23
Timing...4.24
Nature of the Tasks 4.24
Teacher Preferences about Feedback...4.25

Cost Analysis 4.25

SSI Cost Estimates...4.25

Technical Quality..4.26

Development...4.26
Reliability...4.26
Validity 4.28

Conclusions and Recommendations 4.28

Administration of Assessment...4.28
Scoring 4.28
Assessment Content...4.29
Assessment Format 4.30
Summary..4.30

CHAPTER 5: SEMI-STRUCTURED INTERVIEW: ELEMENTARY MATHEMATICS 5.1

Administration of Assessment 5.2

Overview 5.2
Logistics 5.2
Security 5.4
Assessors and Their Training 5.5
Teacher and Assessor Impressions of Administration 5.6

Scoring 5.6

Scoring Process 5.7
Discussion of Scoring System 5.7
Scorers and Their Training 5.8
Teacher Preferences about Feedback 5.9

Assessment Content 5.9

Congruence with California Curriculum Guides and Frameworks 5.10
Extent of Coverage of California Standards for Beginning Teachers 5.11
Job-relatedness 5.15
Appropriateness for Beginning Teachers 5.15

Perceptions 5.17
Performance on assessment tasks 5.18
Comparison of beginning and experienced teachers 5.22

Appropriateness across Contexts 5.23

Grade level and subject matter 5.23
Diverse students 5.24

Fairness across Groups of Teachers 5.24
Appropriateness as a Method of Assessment 5.25

Assessment Format 5.27

Clarity of Assessment 5.27
Format Features 5.29

Timing 5.29
Choice of tasks 5.29
Use of probes 5.31
Use of interviewers 5.31

Cost Analysis 5.33

Technical Quality 5.33

  Development 5.33
  Reliability 5.34
  Validity 5.34

Conclusions and Recommendations 5.34

  Administration of Assessment 5.34
  Scoring 5.35
  Assessment Content 5.36
  Assessment Format 5.37
  Summary 5.37

CHAPTER 6: ELEMENTARY EDUCATION EXAMINATION 6.1

Administration of Assessment 6.2

  Overview 6.2
  Logistics 6.2
  Security 6.4
  Assessors 6.5
  Scoring 6.5
  Teacher, FWL, and RMC Staff Impressions of Administration 6.5

Assessment Content 6.7

  Congruence with California Curriculum Guides and Frameworks 6.8
  Extent of Coverage of California Standards for Beginning Teachers 6.10
  Job-relatedness 6.12
  Appropriateness for Beginning Teachers 6.14
  Appropriateness across Contexts 6.15

    Grade level and subject area 6.15
    Diverse students 6.16

  Fairness across Groups of Teachers 6.16
  Appropriateness as Method of Assessment 6.16
  Comparison with Other Multiple-Choice Tests 6.18

Assessment Format 6.18

  Clarity of Oral Overview and Directions 6.18
  Clarity of Items 6.19
  Timing of Tests 6.19
  Feedback 6.20

Cost Analysis 6.20

Technical Quality 6.21

Conclusions and Recommendations 6.21

    Administration of Assessment 6.21
    Assessment Content 6.22
    Assessment Format 6.23
    Summary 6.24

CHAPTER 7: SEMI-STRUCTURED INTERVIEW: SECONDARY SOCIAL SCIENCE 7.1

CHAPTER 8: CONCLUSIONS 8.1

Assessment Approaches 8.1

    Classroom Observation 8.1

        Definition 8.1
        Characteristics of instruments piloted 8.2
        Strengths and weaknesses 8.2

    Semi-Structured Interviews 8.3

        Definition 8.3
        Characteristics of instruments piloted 8.3
        Strengths and weaknesses 8.3

    Multiple-Choice Examinations 8.4

        Definition 8.4
        Characteristics of instruments piloted 8.4
        Strengths and weaknesses 8.4

    Framework for Comparing Differing Assessment Approaches 8.4

    Costs Estimates 8.6

Next Steps in Designing a System of Teacher Assessments 8.8

    Future Pilot Tests 8.8
    Issues in Design of an Assessment System 8.10

TABLES

TABLE 3.1 Pilot Test Participants: Connecticut Competency Instrument (CCI) 3.6

TABLE 3.2 Connecticut Competency Instrument: Subjects by Grade Levels Observed 3.7

TABLE 3.3 Congruence of CCI with California Curriculum Guides and Frameworks 3.14

TABLE 3.4 Extent of Coverage by the CCI of California Standards for Beginning Teachers 3.18

TABLE 4.1 Semi-Structured Interview in Secondary Mathematics: Pilot Test Participants 4.4

TABLE 4.2 Coverage of the California Mathematics Framework by SSI-SM 4.16

TABLE 4.3 Extent of Coverage by the SSI-SM of California Standards for Beginning Teachers 4.20

TABLE 4.4 Percentage of Rater Exact and Adjacent Agreement by Task and Topic Pairs 4.27

TABLE 5.1 Semi-Structured Interview in Elementary Mathematics: Pilot Test Participants 5.3

TABLE 5.2 Coverage of the California Mathematics Framework by SSI-EM 5.12

TABLE 5.3 Extent of Coverage by the SSI-EM of California Standards for Beginning Teachers 5.16

TABLE 5.4 Semi-Structured Interview in Elementary Mathematics: Scoring Results 5.19

TABLE 6.1 Elementary Education Examination: Pilot Test Participants 6.3

TABLE 6.2 Elementary Education Examination Spring 1989 Pilot Test Results 6.6

TABLE 6.3 Extent of Coverage by the Elementary Education Examination of California Standards for Beginning Teachers 6.13

TABLE 8.1 Analysis of Alternative Assessment Approaches and Their Ability to Assess Specific Teaching Competencies 8.7

TABLE 1 Mean Item Ratings, Coefficient Alphas, and Inter-Rater Correlations for Tests Based on Indicators A.2

TABLE 2 Mean Item Ratings, Coefficient Alphas, and Inter-Rater Correlations for Tests Based on Consensus Indicators A.3

TABLE 3 Mean Item Ratings, Coefficient Alphas, and Inter-Rater Correlations for Tests Based on Indicators, Rater Combinations Reversed A.4

TABLE 4 Adjusted Coefficient Alphas Based on Indicators, Estimated for Equal Test Lengths of 10 Items A.6

TABLE 5 Mean Item P-Values, Coefficient Alphas, and Inter-Rater Correlations for Tests Based on Dichotomized Indicators A.7

CHARTS

CHART 3.1 Defining Attributes for Two Indicators: Connecticut Competency Instrument 3.3

CHART 3.2 CCI Scoring Results 3.21

FIGURES

FIGURE 1 Elementary Education Exam Sample: California Pilot Test Analysis Sample, Social Sciences Majors (N=45) C.3

FIGURE 2 Elementary Education Exam Sample: California Pilot Test Analysis Sample, Science Majors (N=13) C.4

FIGURE 3 Elementary Education Exam Sample: California Pilot Test Analysis Sample, Liberal Arts Majors (N=283) C.5

FIGURE 4 Elementary Education Exam Sample: California Pilot Test Analysis Sample, Education Majors (N=20) C.6

APPENDICES

APPENDIX A SSI-SM Reliability A.1

APPENDIX B SSI-SM Scoring Materials B.1

APPENDIX C Elementary Education Exam Construct Validity C.1

CHAPTER 1:
INTRODUCTION

CHAPTER 1:

INTRODUCTION

Throughout the nation there is renewed interest in and commitment to educational excellence as shown by the many recent analyses of American education (Boyer, 1983; Goodlad, 1984; President's Commission for Excellence in Education, 1983) and the proposals that have been made for educational reform (Holmes Group, 1986; Shulman, 1987; Carnegie Corporation, 1986). Although many different aspects of the educational enterprise have received attention and suggestions for improvement, there has been particular emphasis on the preparation, support, and credentialing of new teachers. This emphasis on new teachers has been part of a broader discussion of the further development of teaching as a profession (e.g., Wise and Darling-Hammond, 1987; Shulman and Sykes, 1986), principally through an increased emphasis on improved quality, opportunities for professional development, and expansion of career roles for classroom teachers.

Policy efforts to increase the quality of teachers have concentrated on improved methods of assessing teacher performance, particularly among beginning teachers. Several leading advocates of educational reform have argued that more rigorous and comprehensive assessments of teachers' knowledge and competence should be developed and adopted (Holmes Group, 1986; Shulman, 1987; Carnegie Corporation, 1986). Of course, efforts to improve teacher quality must also be concerned with maintaining a sufficient quantity of teachers. In California, more than 25,000 teacher candidates were enrolled in collegiate training programs during 1988-89, costing the state hundreds of millions of dollars annually (Gomez, 1989). Unfortunately, up to half of the state's beginning teachers leave their classrooms within five years. This high rate of attrition compounds the recruitment problems of school districts, and increases the overall cost of preparing a sufficient supply of new teachers.

While some new teachers leave the profession to earn higher incomes in other jobs, growing evidence suggests that the high turnover rate among new teachers is also due to a lack of support during the beginning years of teaching. Many new teachers quit because of frustration, isolation, and a sense of inability to meet the increasingly complex demands that all teachers face.

Like most other states, California has many programs for the preparation and certification of prospective teachers. Most of these programs are offered by accredited colleges and universities; some are administered by local school districts, often in conjunction with post-secondary institutions. In addition, many California teachers are trained in other states. Each teacher preparation program in California is evaluated periodically by the Commission on Teacher Credentialing, on the basis of program quality standards, which are designed to ensure that each candidate has a thorough and effective preparation for classroom teaching. Nevertheless, some legislators and other policy makers have advocated a more "candidate-based" certification system, in which the competence and performance of each candidate would be measured and verified in a standardized process. Some teacher advocates have supported the same concept in order to add to the professional stature of teaching. Several universities have advocated "candidate-based" assessment as a way to replace or reduce the evaluation of campus-based programs of teacher preparation. Other teachers and teacher educators have expressed concerns about the potential effects of a standardized assessment process on the attractiveness of teaching as a career, on the diverse composition of teaching as a profession, and on the curriculum of teacher preparation. Many policy analysts have questioned the costs of a statewide assessment process, and researchers have wondered whether valid, reliable measurements could be done at a relatively low cost level.

During the last several years, education policymakers in California have discussed these teacher induction issues, and are interested in examining the extent to which a state policy to support and assess new teachers would:

- improve the effectiveness of new teachers;
- increase the retention of capable new teachers in the profession;
- improve the process of screening teachers' competence as a basis for certification;
- promote professionalism and commitment to professional development among teachers; and
- contribute to school improvement through greater collegiality and involvement in induction by experienced teachers.

Before describing the pilot study of policy alternatives, the relevant literature will be summarized to indicate what is currently known about the characteristics of new teachers which might guide their support and assessment.

Review of Literature on New Teachers

Although education is usually characterized by diverse opinions and controversy, in the past few years a general consensus has been reached regarding the role and importance of new teacher support programs in improving the education of youth in the United States. The consensus is that such programs should seek to address two major concerns: the incompleteness of preservice training and the high departure rates of new teachers from the profession.

Incompleteness of Preservice Training

Effective teachers have mastery of basic concepts in a number of different fields, including human development, psychology, sociology, philosophy, communication, and the disciplines underlying the subjects they teach. They are also familiar with current instructional technology, theories of cognition, and principles of human motivation. Effective teachers quickly grasp the philosophy of a school district's curricular goals, and translate these goals into classroom instructional activities. Furthermore, they know how to adapt instructional strategies to the needs of a variety of learners.

Clearly, the knowledge base of teaching is very complex. Other professions such as medicine, engineering and architecture require a lengthy training period with gradual increases in responsibilities (Wise and Darling-Hammond, 1987). Under the current structure of teacher preparation, prospective teachers are supervised for a brief period of time as they practice applying the concepts and techniques they have learned. They then assume full responsibility for teaching their own classes with minimal supervision and support. Although collegiate study and supervised student teaching are important rehearsals, they do not represent the many complex and varied situations a new teacher faces in his or her own classroom. The literature on new teachers shows a growing realization that support programs for beginning teachers are needed to complete the training and studies that prospective teachers experience in colleges and universities (Borko, 1986; Clark et al.; McDonald, 1980). When new California teachers in one study commented on their teacher preparation experience, they repeatedly noted their difficulty in applying what they learned in coursework to their present classrooms (Berliner et al., 1987). For this reason, teacher educators have recognized for many years that preservice courses and experiences cannot fully prepare college students to perform as excellent practitioners in classrooms. Regardless of the quality and effectiveness of prior study and supervised practice, the acquisition of professional practices in contemporary classrooms requires extended opportunities to reflect on and discuss these practices in a collegial environment.

Problems of New Teachers

New teachers face problems in three areas: technical, socioemotional, and institutional. Technical problems are related to content transmission, pedagogy and management of the classroom. It is common for new teachers to report significant difficulties in classroom management (Veenman, 1984), curriculum implementation (Grant and Zeichner, 1981; Veenman, 1984; Berliner et al., 1987), managing diversity within the classroom (Grant and Zeichner, 1981; Veenman, 1984; Borko et al., 1986; Berliner et al., 1987; Berliner et al., 1988), motivation of students (Veenman, 1984) and relations with parents (Veenman, 1984; Berliner et al., 1987). The few extant studies of the effectiveness of teacher support projects (Varah et al., 1986; Huling-Austin, 1988) provide some evidence that induction projects can affect the instructional effectiveness of new teachers in comparison with either a control group of teachers who did not receive formal support, or the new teachers' effectiveness at the beginning of the support project.

In addition to technical problems, new teachers also report socioemotional problems. Most teachers work in isolation from other adults, with few opportunities to observe their colleagues (Lortie, 1975). Consequently, they have few chances to compare their classes and teaching to that of other teachers, or to determine how their problems and successes compare with those of other teachers. This lack of information and uncertainty magnifies the insecurity and self-doubts of new teachers, who face the problems of acquiring and developing materials, lesson plans and tests without the experience and expertise that seasoned teachers draw upon. At times, the many demands of the classroom intrude on new teachers' personal lives. Not surprisingly, they generally appreciate someone who is willing to listen to their problems -- both personal and professional -- and offer supportive and useful feedback (Borko, 1986).

The last category of problems that new teachers face are institutional ones. These include the task of understanding district and school policies, practices, and procedures; identifying resources and how to take advantage of them; and becoming a member of the community of teachers in the school. Many new teachers have initial difficulties in locating and absorbing this critical information (Grant and Zeichner, 1981; Odell, 1986).

Differences between Novice Teachers and Expert Teachers

As the induction of teachers into the profession has emerged as a major issue in education, researchers have begun to study the differences between novice and expert teachers. Although we are only in the initial stages of understanding the development of expertise in teaching, and in identifying the extent of variation among individual teachers, this knowledge is particularly critical to the design of support programs and assessments of beginning teachers. As the process of teacher development is better understood, it is likely that the principles which guide the practices of expert teachers can be incorporated into new teacher support programs. If so, perhaps larger numbers of beginning teachers would attain higher levels of expertise and effectiveness, which would allow more in-depth assessment of particular knowledge and skills. This section summarizes briefly the literature on differences between novice and expert teachers.

In comparison with new teachers, when experienced teachers are asked to describe their own lesson plans (Leinhardt, 1989) or what they observe in a videotape of another classroom's activities (Berliner, 1989), the experienced teachers provide more detailed descriptions, and their descriptions exhibit more cohesive themes. Experienced teachers seem to see lessons as composed of general pedagogical routines--routines for introducing new concepts, routines for applying concepts previously learned, routines for reviewing material previously learned, homework collection routines, and groupwork routines. Novice teachers are unlikely to use the language of routines to describe classroom activities; indeed, they do not seem to perceive the importance of routines. The time spent in routine activities is much more variable for novice teachers than for experienced teachers; novice teachers also use a more varied and loosely coupled set of activities than experienced teachers. The result is that they frequently need to spend time familiarizing students with and enforcing rules regarding the activity, reducing efficiency in the use of classroom time (Leinhardt, 1989).

This variation in general pedagogical skills is paralleled by variation in the skills of content pedagogy. There is less variation in representations of content used by experienced teachers than in those used by novice teachers. Experienced teachers are more likely to use the same representation of content (e.g., a number line) for a series of lessons, while novice teachers often use unfamiliar representations of content to introduce new concepts (Leinhardt, 1989). The practice of the novice teachers results in more confusion by students who needed to learn a new form of content as well as a new concept. The explanations of concepts by experienced teachers were also more concise, highlighting prerequisite skills or concepts that the students already had learned (Leinhardt, 1989).

The content knowledge of expert and novice teachers also differs in a similar fashion. Expert teachers see the subject as organized in frameworks; novice teachers are more likely to see it as a collection of facts (Wilson, 1988; cf. Leinhardt, 1989). The extent to which this organization of content knowledge is likely to develop with classroom experience is not clear. It seems plausible that a certain level of content knowledge is prerequisite to conceptualizing the subject matter in terms of frameworks.

Expert teachers also are better able to articulate a knowledge of students and student learning than either novice teachers (Leinhardt, 1983; Wilson, 1988) or subject matter specialists whose major focus is content and not teaching (Wilson, 1988). This includes generic knowledge, subject-specific knowledge, and topic-specific knowledge about students, which experienced teachers formulate into frameworks which guide both general and individual instruction.

These recent studies of novice and expert teachers suggest there is much that new teachers need to learn in order to become proficient classroom practitioners. As was suggested previously, however, many individuals cannot learn all of the complexities of teaching in preservice training programs or supervised practicums. An extended process of intensive consultation and mentoring is needed for beginning teachers to acquire the skills and knowledge that constitute expertise in pedagogy.

California New Teacher Project

To explore innovative methods of new teacher support and assessment, the California Legislature, in the Teacher Credentialing Law of 1988 (Chapter 1355 of the Statutes of 1988), created the California New Teacher Project (CNTP). The CNTP, jointly administered by the Commission on Teacher Credentialing (CTC) and the State Department of Education (SDE), has three components: support, evaluation, and assessment. A brief overview of each component and the overall goals of the CNTP are found in this section; the assessment component is described in more detail in the following section.

During the first year (1988-89), the support component of the CNTP consisted of fifteen local pilot projects representing diverse teaching contexts as well as a variety of approaches to supporting new teachers. Approximately 650 first- and second-year teachers participated in training or seminars sponsored by districts and institutions of higher education, worked with mentors and other experienced teachers, and met with peer support groups. Support projects funded by the California New Teacher Project differ in their areas of emphasis and methods of delivery, but they collectively address the technical, socioemotional, and institutional problems of new teachers. It should be noted that these projects are not the only new teacher support programs in California; others are sponsored by individual school districts or jointly by the SDE and the California State University system. However, these fifteen projects have agreed to participate in research on alternative methods of new teacher support that is sponsored by the CNTP.

The CNTP evaluation component is investigating the effects of the support on new teacher effectiveness and retention, as well as the cost-effectiveness of the various methods used to support new teachers in the fifteen projects. Experimentation with alternative methods of teacher support combined with the evaluation of these forms of support should help to identify the kinds of assistance that are most effective as new teachers enter the profession. The CTC and SDE have contracted with the Southwest Regional Laboratory (SWRL) to conduct all activities in the evaluation component. The results of this evaluation are being reported to the CTC and SDE in a separate report by SWRL.

Assessment Component of the California New Teacher Project

The assessment component of the CNTP consists of the development and pilot testing of innovative forms of new teacher assessment. The evaluation of diverse approaches to teacher assessments is intended to identify the most promising ways in which a comprehensive assessment of teacher candidates could inform the certification process and contribute to the quality of teaching. This document reports the analysis of the pilot tests that were conducted in the assessment component in 1989, the first year of the CNTP. The pilot tests were administered by Far West Laboratory for Educational Research and Development (FWL), with assistance in the design, observation, and analysis of the pilot tests from RMC Research Corporation. The design and purpose of these pilot tests are described in Chapter 2. During 1989, the assessment component also included the development of five additional assessments that will be pilot tested in the second year of the project.

The Bergeson Act (S.B. 148) specifically requires that each alternative method of support and assessment be evaluated in terms of:

- its effectiveness at retaining capable beginning teachers in the profession;
- its effectiveness at improving the pedagogical content knowledge and skills of the beginning teachers who are retained;
- its effectiveness at improving the ability of beginning teachers to teach students who are ethnically, culturally, economically, academically, and linguistically diverse;
- its effectiveness at identifying beginning teachers who need additional assistance and, if that additional assistance fails, who should be removed from the profession of education;
- the relative costs of the method in relation to its beneficial effects; and
- the extent to which an alternative method of supporting or assessing beginning teachers would, if it were added to the other state requirements for teaching credentials, make careers in education more or less appealing to prospective teachers.

Although the support and assessment components are guided by relevant state curriculum frameworks and expectations for the pedagogical competence of new teachers, the SDE and CTC have not generated a list of competencies to serve as a common focus for all components of the CNTP. Instead, to increase the variety of methods being evaluated, the assessment component is conducted independently of the evaluation and support components. For this reason, the competencies being measured by the assessment instruments being piloted may or may not coincide with the areas of support offered to the new teachers by their support projects. The integration of lessons learned from the evaluation and assessment components will facilitate an analysis of relationships and interactions among teacher preparation, support, assessment and certification to suggest whether and how a program of support and assessment for new teachers should be developed in a coordinated manner.

Since the purpose of this document is to describe and analyze the pilot testing for the assessment component, the rationale and design for this component are described in more detail in the following section.

Rationale and Design of the Assessment Component

As a result of educational reform efforts which focus on the development of teaching as a profession, states are moving to candidate-based assessments to supplement their existing program-based modes of assessment. To enhance the academic abilities of teacher candidates, there is a trend toward setting higher standards for teacher preparation programs, increasing the requirements for matriculation, and specifying the competencies to be mastered before the completion of programs. States are also adopting new assessments whose passage by teacher candidates is required for credentialing.

Like other states, California is particularly concerned about maximizing the quality of teaching in its schools. The CTC has recently revised the Standards of Program Quality and Effectiveness (CTC, 1988) for teacher preparation programs. The Commission's new standards include a definition of the levels of pedagogical competence and performance expected of program graduates. California has also participated in the movement to evaluate individual teacher candidates through the use of particular instruments that assess teacher competence: the California Basic Educational Skills Test (CBEST), the NTE Core Battery, and the NTE Specialty Area Tests. In recent years, these tests have been reviewed by California teachers and teacher educators in terms of their appropriateness for use in the credentialing process (Watkins, 1985; Wheeler and Elias, 1983; Wheeler et al., 1988). Suggested changes in the fifteen NTE Specialty Area Tests, for example, included revision of the content to make the tests more compatible with the California Curriculum Frameworks, and augmentation of these multiple-choice tests with some type of performance assessment. These changes are currently being implemented by the CTC in consultation with the State Superintendent and Educational Testing Service.

Nationally, the interest in assessing the quality of teachers has underscored the absence of assessment approaches that are closely related to the tasks that teachers perform in the course of their work. This has led to the development of alternatives to multiple-choice tests, which historically have been the dominant form of large-scale teacher assessments. The alternatives are often referred to as "innovative" or performance-based assessments because of their emphasis on direct measurement of actual teacher performance.

A variety of performance-based teacher assessments has been developed in recent years, including a number of observation instruments which have been adopted as requirements in teacher certification programs in other states. However, many of these instruments are quite prescriptive in terms of teaching style. Most of them simply measure the frequency of specific behaviors that are generally associated with student achievement, rather than assessing the appropriateness of such behaviors when they occur in particular situations. The Bergeson Act specifically prohibits the use of checklists of teacher behaviors which tabulate the presence or absence of discrete behaviors. Since California classrooms are extremely diverse, instruments which do not fairly assess a variety of teaching styles in diverse contexts are inappropriate for use in assessing California teachers. For this reason, the CNTP is designed to evaluate the degree to which various assessment approaches measure the ability of teacher candidates to teach a wide variety of students.

The Bergeson Act reflects an emerging design for California's assessment of teacher candidates in four areas: basic academic skills; subject matter knowledge; subject specific pedagogy; and general pedagogy. The CBEST has been judged to be suitable for assessing candidate performance in the first area (Watkins, 1986). Two current projects that address the second area are: the development of a replacement test for the NTE Core Battery Test for the assessment of subject matter knowledge of prospective elementary teachers; and the revision and augmentation of the NTE Specialty Area Tests for assessing subject matter knowledge of prospective secondary teachers. The last two areas were judged to be best assessed after teacher candidates have some experience in conducting their own classrooms, i.e., in the first year or two of teaching. The CNTP focuses on the identification of promising, cost-effective methods of assessment of the last two areas, with special emphasis on six content areas: Elementary Teaching, Secondary English, Secondary Mathematics, Secondary Life Science, Secondary Physical Science, and Secondary Social Science.

Because of the high interest in teacher assessment among education professionals in recent years, together with a growing recognition of the limitations of multiple-choice forms of assessment, new assessment approaches are being developed. These new approaches include the evaluation of tasks that resemble those which teachers commonly perform in the classroom. Assessment approaches include the use of videotapes or written vignettes, structured interviews, structured simulations, and reviews of portfolios of a teacher's work. Classroom observation instruments, which assess teachers in the course of instruction in their own classrooms, are also being revised and refined.

In planning the research to be conducted in the assessment component of the CNTP, staff from the CTC and SDE considered both the high cost of assessment development and the desirability of evaluating a wide variety of assessment approaches. Many "innovative" assessment instruments are in the initial stages of development, and could serve only as initial prototypes for exploring the potential of an assessment approach, rather than as state-of-the-art instruments reflecting a long period of experimentation within the approach.

To maximize the information to be gathered while minimizing the costs, the instruments chosen for pilot testing in the initial year (and the instruments to be developed in the subsequent years) were not required to be fully developed products whose validity and reliability were well established. Instead, the pilot testing was designed to yield information about the strengths and weaknesses of assessment ***approaches*** for which the specific instruments serve as outstanding exemplars. The purpose of the pilot testing is not to consider particular instruments for adoption, but to identify promising approaches to the assessment of teachers, to guide future selection and/or development of assessment instruments which are tailored to the California context. Consistent with this purpose, assessment prototypes were piloted on a small scale with a thorough trouble-shooting process in order to learn as much as possible about the effects of each approach before conducting expensive, large-scale field tests.

To ensure the broadest possible representation of assessment approaches, CTC and SDE staff began with a review of existing teacher assessment instruments. They hoped to avoid as much as possible the high costs of initial development by pilot testing existing instruments. However, the state agencies were able to locate and identify only a few existing assessment instruments that employ innovative modes of assessment, or that assess significant domains of teacher competence that have not been assessed adequately in the past. These instruments include: a classroom observation instrument that assesses general pedagogy; three semi-structured interviews in elementary mathematics, secondary mathematics, and secondary social studies which assess subject-specific pedagogy; and a multiple-choice examination that utilizes innovative questions and materials to assess general pedagogy and content-specific pedagogy in elementary education. The CTC/SDE staff chose to pilot test these instruments in the initial year of the CNTP, and to commission the development of additional instruments in other areas which had been insufficiently explored. These additional instruments are slated for pilot testing in the second and third years of the project.

A comprehensive teacher assessment system for California cannot be developed quickly. For example, classroom observation instruments should reflect the complexities of student-teacher interactions, instructional decisions, and student involvement. Most performance-based assessments that would capture these complexities are in initial stages of development, and would need to be tailored to the California curriculum and diverse teaching contexts. During the next two years, the experimental work to be undertaken by the CNTP will provide insights into the kinds of assessments that would be most cost-effective, when and how those assessments should be administered, and how educational groups and organizations can best assist prospective and novice teachers in environments that feature assistance and accountability. Although the main purpose of the assessment component is to evaluate assessment approaches for use in credentialing teacher candidates, their capacity to advise teacher candidates of their strengths and weaknesses, and to guide the choice of staff development or induction activities, will also be considered.

The specific contributions of the Spring 1989 round of pilot testing are discussed in the following section.

Pilot Testing in 1989

The purpose of the pilot testing in 1989 was to examine in California the functioning of several assessment instruments which are considered to be promising exemplars of innovative assessment approaches. The evaluation of the various components (e.g., logistical requirements, prompt materials, scoring criteria, training exercises for assessors) of these instruments was intended to provide information concerning the strengths and limitations of the assessment approaches which the specific instruments represented. The pilot tests were not expected to yield definitive measurements of the psychometric properties of the instruments because the prototypes had not been sufficiently developed for that to occur. This focus on trouble-shooting allows small-scale pilot testing, requires fewer resources, and considerably increases the number of assessment approaches which can be examined. The goal of the pilot tests is to suggest whether or not it is advisable to invest additional resources in the development of assessments resembling those piloted.

This document is the final report and analysis of the Spring 1989 pilot test administration. Each of the assessment instruments is described, and the ease of administration, scoring, content and format, costs, and technical qualities are analyzed. The administration of the pilot tests was described in detail in the *Administration Report for Spring, 1989.*

Terminology Used in the Report

Specialized terms and abbreviations which appear in this report are defined below:

Assessment: the process of measuring the performances of new teachers in order to help them improve, and to determine whether their performances satisfy one or more standards for professional certification as classroom teachers.

Assessor: the person who administers an assessment instrument.

Candidate: a person participating in an assessment for the purpose of satisfying requirements for earning a teaching credential.

CCI: Connecticut Competency Instrument. A classroom observation instrument developed by the Connecticut State Department of Education.

CNTP: the California New Teacher Project, which evaluates methods of new teacher support and assessment. The project has three components: sponsorship of new teacher support projects (which numbered 15 at the time of the Spring 1989 pilot testing); evaluation of various methods of teacher support exemplified by these support projects; and pilot testing of innovative assessments of new teachers.

CTC: the California Commission on Teacher Credentialing. The CTC staff shares responsibility for overseeing the California New Teacher Project.

FWL: Far West Laboratory for Educational Research and Development. FWL is administering the assessment portion of the California New Teacher Project and analyzing the potential of alternative assessment approaches for possible future use as new credentialing requirements.

IOX Assessment Associates: developers of the Elementary Education Examination which was piloted as an example of an innovative form of the multiple-choice test approach to assessment.

Project: one of the fifteen support projects in the California New Teacher Project sponsored by the CTC/SDE.

Project Director: a director of one of the fifteen projects.

SDE: the California State Department of Education. The SDE staff administers the California New Teacher Project jointly with the CTC. Often the two will be referred to jointly as the CTC/SDE.

RMC: RMC Research Corporation. RMC staff are collaborating with FWL staff in the design and analysis of the pilot tests.

SWRL: Southwest Regional Laboratory. SWRL is conducting the evaluation of new teacher support methods exemplified by the CNTP projects.

TAP: the Stanford Teacher Assessment Project, which develops prototypes of assessments to be used to certify expert teachers.

Teacher: first- and second-year teachers with California teaching credentials.

The next chapter describes the pilot test design and the processes used to evaluate the assessment approaches which were examined in the spring of 1989. The next chapters discuss the pilot tests of specific instruments in the following order: the Connecticut Competency Instrument (CCI), the Semi-Structured Interview in Secondary Mathematics, the Semi-Structured Interview in Elementary Mathematics, and the Elementary Education Examination. Reasons for postponing the pilot test of the Semi-Structured Interview in Secondary Social Science are also discussed. The report concludes with a summary of general lessons learned about performance assessments and recommendations for next steps.

CHAPTER 2:
PILOT TEST DESIGN AND ANALYSIS

CHAPTER 2:

PILOT TEST DESIGN AND ANALYSIS

This chapter describes the design and analysis of the pilot tests of prototypes representing various assessment approaches. Subtopics include the source of instrumentation, the sampling plans, sources of information for evaluating the instruments and the assessment approaches, methods of data reduction and major categories of analysis. Deviations from the design due to unanticipated events will be described in following chapters which focus on the individual instruments. The analysis of the pilot tests is contained in two reports. The first report, *Administration Report for Spring, 1989*, describes the administrative aspects of the pilot tests of the different assessment instruments and discusses teacher responses. This final report focuses more on the content and evaluation of the prototypes, and recommends next steps for the pilot testing of additional prototypes.

Design of Pilot Tests

This section on the design of the pilot tests describes the sources of instruments and the sampling plans. Procedures for data collection and analysis will be described in the section on analysis of the pilot tests.

Sources of Instrumentation

The four instruments were selected after an extensive search by CTC and SDE staff for promising prototypes of innovative assessment formats. The sources of the instruments varied, so each will be described separately.

The Connecticut Competency Instrument (CCI), a classroom observation instrument that measures general teaching effectiveness in elementary and secondary schools, was developed by the Connecticut State Department of Education. Observers who had previously been trained in the use of the CCI by the State of Connecticut were used to conduct the observations.

The semi-structured interviews came from two different sources. The Semi-Structured Interview in Secondary Mathematics (SSI-SM) was developed by the State of Connecticut, which provided previously trained assessors from Connecticut to administer the assessment. The scoring system was in the process of development at the time of administration; substantial progress in development was made, and portions of the interviews were scored. The Semi-Structured Interviews in Elementary Mathematics and in Secondary Social Studies (SSI-EM and SSI-SSS respectively) were developed by the Teacher Assessment Project (TAP) at Stanford University as part of their work with the National Board for Professional Teaching Standards. The interviews from Stanford were originally developed to identify expert teachers, so the questions and scoring system were revised to be more appropriate for beginning teachers. No trained assessors were available, so a TAP representative trained the assessors and scorers for all Stanford instruments.

The Elementary Education Examination, a multiple-choice test, was designed for beginning teachers by IOX Assessment Associates (formerly called the Instructional Objectives Exchange) under a contract with the State of Connecticut. Although we refer to it as a "test," it is actually a collection of items placed into six test forms. These items were pilot tested to assess their feasibility for incorporation into a test of competence in elementary education which includes both pedagogy and content knowledge. IOX provided all materials and assumed full responsibility for administering and scoring the items.

Sampling Plan

Several factors constrained the construction of the pilot test sampling plan. The first was the necessity of planning, scheduling and administering five assessments within a three-month period. The second was the desirability of using the teachers in the fifteen pilot projects who had already consented to participate in assessment pilot testing.

We began the sample selection process by assembling lists of possible participants within each project. Once these lists were completed, the characteristics of grade level, school context (urban, suburban and rural), gender and ethnicity were considered in selecting teachers from those projects with a suitable concentration of teachers with the appropriate credential. (The threshold number varied with the particular assessment instrument being piloted, ranging from eight for the semi-structured interviews to thirty for the multiple-choice examination.) For example, for the SSI-EM, lists of secondary math teachers were assembled, and projects with more than eight teachers were contacted. Although we wanted to maximize variation in the characteristics of teachers selected, our ability to do so was limited by the information which we had about project teachers, the time required to recruit nonproject teachers, and the small samples. Information on the ethnicity of teachers was available for many of the projects, but there were few nonwhite teachers, precluding the selection of a significantly large subsample. Our information on school context was minimal, based solely on classifications of districts provided by the New Teacher Support Projects. We also tried to include teachers from each project in at least one pilot test, though no attempt was made to equalize the participation rate across projects.

Considerations of administration costs and time constraints led us to not include some project teachers from remote areas. In the case of the classroom observation assessment and one of the semi-structured interview assessments, the use of Connecticut assessors who were only available for a specific week limited flexibility in selecting teachers because of constraints on time for travel to multiple sites. To complete the pilot testing on schedule, the recruitment of nonproject teachers was limited to those with the appropriate credential who were located near an identified sample of project teachers. Most nonproject districts could identify teachers in their first year in the district, but could not readily determine whether these teachers were in their first year of teaching. Some nonproject districts contacted had a time-consuming approval process required for the release of teachers' names. Therefore, the use of nonproject teachers was minimized.

The characteristics of teachers in the samples are described in more detail in the chapters that focus on specific instruments.

Analysis of Pilot Tests

This section describes our procedures for data collection and reduction, as well as the key analytic categories focusing on specific aspects of instruments. The data collected also served as a basis for judging the potential of the assessment approach which the particular instrument utilized.

Data Collection

Since the same means of data collection were used for all assessment instruments, they will be discussed together. Several sources of data were used:

- o evaluation feedback forms for teachers who participated in the pilot tests;
- o evaluation feedback forms for the assessors and scorers;
- o observations of the administration of each assessment and the training of assessors and scorers recorded in field notes by FWL and RMC staff;
- o scores that reflected the performances of participating teachers on the assessment instruments; and
- o the relevant Curriculum Guide(s) or Framework(s) and the California Standards for Beginning Teachers.

Following guidance from the funding agencies, RMC staff developed an outline of issues to be addressed in evaluation feedback forms to be completed by participating teachers, assessors, and scorers. FWL staff then developed separate forms for each group which were tailored to specific assessment instruments. These forms were given to teachers upon the completion of each assessment, except in the case of the classroom observation instrument, where they were mailed. Assessors and scorers returned completed forms when they presented invoices for payment. Since the emphasis in the pilot tests was on trouble shooting, the evaluation feedback forms focused on critical evaluations of the instruments with respect to the analytic categories described later. Most of the questions were either open-ended or required yes/no answers with spaces provided to elaborate.

Field notes were taken during observations of the assessment administrations. FWL and RMC staff attended most administrations of the assessment instruments. FWL staff observed the training of assessors and scorers. When they were familiar with the subject matter, FWL staff also served as participant observers for scoring to obtain a more complete understanding of the performance of the assessment instruments. RMC staff served as participant observers for the classroom observation instrument. FWL staff also participated as assessors for some administrations of one of the semi-structured interviews.

All of the instruments were scored; the interpretation of results was not always straightforward because scoring systems varied in terms of the stage or level of development. For example, the classroom observation instrument had a well-developed scoring system which had been previously piloted and revised to produce greater inter-rater reliability. In contrast, the scoring systems for the semi-structured interviews were devised or revised after administration of the pilot tests, and hence were unknown to the interviewers, creating some inconsistencies between questions and/or probes and the scoring categories. Training for scoring varied according to the level of development of the instrument.

The content of each prototype was compared to all of the relevant California Model Curriculum Guides and Frameworks, and with the California Standards for Beginning Teachers. The Model Curriculum Guides and Frameworks are recent documents produced by subject matter panels convened by the California State Department of Education. Reflecting a consensus among panel members on the content and philosophy of instruction, these documents are expected to guide curriculum development and instruction in the subject in California public schools. If there were two or more Guides or Frameworks addressing a particular subject area, the most recent one was used.

The California Beginning Teacher Standards are standards that define the level of pedagogical competence and performance that the Commission on Teacher Credentialing expects the graduates of credential programs to attain as a condition for program approval. These standards--Standards 22 through 32--are listed in *Standards of Program Quality and Effectiveness, Factors to Consider and Preconditions in the Evaluation of Professional Teacher Preparation Programs for Multiple and Single Subject Credentials.* (Other standards address more general program requirements; these focus specifically on candidate competencies.) Although these are standards for teacher preparation ***programs*** and not teacher ***candidates***, they identify the knowledge and skills that beginning California teachers are expected to attain.

Data Reduction

Data reduction techniques varied with the data collection method. Fixed-response questions on the evaluation feedback forms completed by teachers participating in the pilot tests, assessors and scorers were tabulated by hand. Most of the questions, however, were open-ended. Surveys were reviewed, and response categories were developed to code the open-ended responses and comments. In addition, responses which either stated a common viewpoint well, or which provided an additional perspective, were culled for possible quotation in the reports. For the fixed-response questions where elaboration was invited, positive responses were less likely to be elaborated than negative ones, so many more negative evaluations were available for quotation than positive ones.

Field notes were reviewed for relevant information that address the analytic categories and were incorporated into the chapters about specific instruments.

For the multiple-choice examination, there was a large enough sample to permit extensive analysis of scores by subgroups. For other assessment instruments, the general distribution of scores was examined; in some cases, the scores of teachers from nonwhite ethnic groups were examined separately. Some exploratory analyses were performed to assess the internal consistency and rater agreement on the secondary math interviews.

The Model Curriculum Guides and Frameworks were examined by FWL staff with the appropriate subject matter background. Their professional judgments were used to draw conclusions about the congruence of the assessment instruments with the relevant Guide or Framework. The reasoning underlying these judgments are described in detail in the chapters on the specific prototypes.

Overview of Analytic Categories

The same general analytic categories were used to appraise all assessment instruments. They included: administration, content, format, cost analysis, and technical quality. These categories and their subcategories are discussed below.

Administration of assessment. This category included consideration of the logistics, security needs, and training of assessors and scorers for the particular assessment instrument. Generally, this category generated information required to estimate administrative requirements and cost projections. The logistics required for administration predict the ease of administration if the assessment approach were to be implemented on a statewide basis. Generally, the more complicated the logistical requirements, the more expensive the assessment is to administer. The needs for security impact not only logistical requirements, but also the frequency with which the instrument must be revised for statewide administration. Consideration of the training of assessors and scorers suggests the degree of difficulty to be anticipated in recruiting people with the required professional expertise, and the time required to prepare personnel to administer and score the particular assessment instrument.

Assessment content. This category addressed the specific instrument's congruence with the relevant Curriculum Guide or Framework, and the extent to which the California Standards for Beginning Teachers were covered. It also included an examination of the content of the assessments along the following dimensions: job-relatedness, appropriateness for beginning teachers, appropriateness across varying teaching contexts, fairness across different groups of teachers, and general appropriateness of the assessment approach represented by the prototype as a method of assessing teachers. Since none of the assessment prototypes was specifically developed for use in California, comparison of the assessment content with the relevant Curriculum Guide and the California Standards for Beginning Teachers was necessary to determine whether the assessment approach was compatible with the instructional philosophy underlying the various California curricula and the competencies specified for teacher candidates. Since one common criticism of teacher assessment instruments is that scores have not been shown to be closely related to specific teaching competencies, job relevance was included as an analytic category. The more closely the assessment tasks resemble the activities that teachers do in the course of their teaching duties, the higher the potential relationship of scores to actual teaching competencies.

Since the CNTP focuses on the assessment of teachers early in their teaching career, it is important to judge the appropriateness of each assessment in terms of performance expectations and perceived difficulty for teachers at this stage of career development. Appropriateness across contexts is particularly important for California, since it has a wide diversity in student populations. The issue of fairness across groups of teachers relates to the potential for bias with regard to any particular group of teachers.

Assessment format. This category included the general clarity of orientation materials and instructions, as well as the identification of important features peculiar to a particular assessment format. In order for the performance of candidates to reflect their true competencies, it is essential that each candidate have clear and accurate expectations of the performance which is expected of them. This is not possible when teachers are uncertain as to what they are being asked to do. This category also covers features which are peculiar to particular assessment formats identified as either problematic or critical to successful implementation of the assessment approach.

Cost analysis. Based on the pilot testing experience, we attempted to project the costs of a statewide administration of an instrument which resembled the prototype tested.

Technical quality. This category discussed the work performed to date in the development of the prototype. Although few data were available to assess the reliability and validity of any instrument, procedures for doing so were recommended.

This chapter has outlined the general design for the Spring 1989 pilot tests in the assessment portion of the California New Teacher Project. The following five chapters discuss each of the assessment instruments: the classroom observation instrument (Connecticut Competency Instrument or CCI), a semi-structured interview in secondary mathematics (SSI-SM), a semi-structured interview in elementary mathematics (SSI-EM), and an innovative multiple-choice test (Elementary Education Examination).

CHAPTER 3:
CONNECTICUT COMPETENCY INSTRUMENT (CCI)

CHAPTER 3:

CONNECTICUT COMPETENCY INSTRUMENT (CCI)

The Connecticut Competency Instrument (CCI) is a classroom observation system developed by Connecticut's State Department of Education. Through this system an observer conducts a 45-60 minute classroom observation, focusing on ten indicators of a teacher's classroom performance. These 10 indicators, grouped in three clusters to represent three major areas of instruction, are as follows:

I. Management of the Classroom Environment

a. Promoting a positive learning environment
b. Maintaining appropriate standards of behavior
c. Engaging students in activities of the lesson
d. Effectively managing routines and transitions

II. Instruction

a. Creating a structure for learning
b. Presenting appropriate lesson content
c. Developing a lesson to promote achievement of lesson objectives
d. Using appropriate questioning techniques
e. Communicating clearly

III. Assessment

a. Monitoring student understanding and adjusting teaching

In addition to the observation which focuses on the above ten indicators, the CCI system includes a pre-assessment information form, completed by the teacher, which informs the observer of the learning objectives, activities, instructional arrangements, and materials associated with the lesson. Next there is a pre-observation interview in which the observer meets with the teacher to review the information included in the pre-assessment information form. Finally, there is a post-observation interview in which the teacher meets briefly with the observer to explain any deviations from the plan that may have occurred during the lesson.

A key feature of the CCI is the analysis and rating process. After scripting what takes place in the classroom as accurately as possible, the observer completes a one-page form for each of the ten indicators. In one column of the form, the observer writes evidence from the script that supports the indicator, and in another column she/he records evidence that does not. The recorded evidence is specifically tailored to one or more of the attributes that define each of the indicators. For example, for the first indicator, "promoting a positive learning environment," the observer records positive and negative (if any) evidence from the script for each of three defining attributes: rapport, communication of expectations for achievement, and physical environment. Each of these attributes is also defined so the observer would record evidence that, for example, the teacher has or has not established rapport by "demonstrating patience, acceptance, empathy and interest in students through positive verbal and non-verbal exchanges." After the careful recording of evidence, the observer then weighs the evidence in both columns in order to rate the teacher's performance on the indicator as either "Acceptable" or "Unacceptable."

For each of the ten indicators, the CCI includes one or more attributes that elaborate on the meaning of the indicators. Chart 3.1 shows the defining attributes for two of the ten indicators. Connecticut has developed operational definitions for all CCI indicators and attributes. These definitions are an important part of the training of CCI observers.

The CCI is not a typical classroom observation system, but could be considered a "state-of-the-art" representative of the classroom observation approach to teacher assessment. Its attention to specific evidence regarding teaching abilities distinguishes the CCI from most other classroom observation instruments which generally tend to be either structured checklists or rating-scale instruments. It is further distinguished from most other classroom observation instruments in that it (1) acknowledges that competent teaching may be manifested in diverse ways, and (2) emphasizes the importance of the professional judgment of trained assessors in making decisions about teacher competence.

To better understand why the CCI was chosen for pilot testing in California, we can compare the CCI to classroom observation instruments that are used in two other states: Florida and Georgia. Two classroom observation instruments -- a Summative Observation Instrument and a Formative Observation Instrument -- are used in Florida's Beginning Teacher Program. Both require an observer to mark in a box whenever a specified behavior is observed. The behaviors (also referred to as indicators) are organized under six domains and are described as dichotomous pairs. For example, the first behavior for the Summative Observation Instrument is the way a teacher begins instruction. The observer evaluates the teacher's initiation of the lesson and selects either the box marked "promptly" or the box marked "delays," with no intermediate evaluation possible. In a sixty-minute period, the observer is to mark the frequency of twenty-one behaviors, all but four of which are described in dichotomous terms. This instrument, therefore, emphasizes the occurrence of specified teaching behaviors with little or no regard for the appropriateness of those behaviors. For example, it may be appropriate for a teacher to delay initiation of a lesson, but this instrument does not allow the observer -- or the teacher -- to make such a judgment.

The state of Georgia also uses two classroom observation instruments, which are collectively known as the Teacher Performance Assessment Instruments (TPAI), in its assessment of beginning teachers. These instruments require an observer to give a rating, using a five-point scale, to each relevant behavior (or indicator) observed. To guide the observer in giving the rating, each indicator is defined by a range of teaching behaviors referred to as descriptors. In the case of some indicators, the descriptors constitute the rating scale; in other instances, the number of descriptors observed is the basis for scoring a teacher's performance. For example, for the indicator, "uses procedures which get learners initially involved in lessons," four descriptors are provided. A rating of "1" would be given if "none of the descriptors is evident," and a rating of "5" if "four of the descriptors are evident." Between the two instruments, a total of 30 indicators are rated during each 60-minute observation.

CHART 3.1

DEFINING ATTRIBUTES FOR TWO INDICATORS: CONNECTICUT COMPETENCY INSTRUMENT

Indicators	Defining Attributes
IC. THE TEACHER ENGAGES THE STUDENTS IN THE ACTIVITIES OF THE LESSON.	*(1) Student engagement:* The beginning teacher engages a clear majority (at least 80 percent) of the students in the instructional activities of the lesson. Engagement is defined as students' involvement in lesson activities consistent with the teacher's expectations or directions. *(2) Re-engagement:* When any students are persistently off-task, the teacher attempts to bring them back on task.
IID. THE TEACHER USES APPROPRIATE QUESTIONING TECHNIQUES.	*(1) Appropriateness to lesson content:* Questions must be related to the content of the lesson and appropriate to the lesson objectives. *(2) Responding to students:* Teachers should respond to student answers or failures to answer. When appropriate, teachers should also build upon student answers to work toward the lesson objectives. *(3) Opportunities for student involvement:* Opportunities for student involvement are provided by appropriate wait time and by addressing questions to a variety of students, encouraging most students to be involved. Teachers should distribute response opportunities to all students. Wait time should be suited to the type of question asked. *(4) Cognitive level: Level of questioning* Level of questioning must be appropriate to the teacher's objectives. If the teacher is seeking recall of basic facts or concepts, then lower-order cognitive questions may be appropriate. If the teacher's purpose is to stimulate higher-order thinking, problem-solving or generalizing, then higher-order cognitive questions should be asked. In many lessons, a variety of questioning levels will be appropriate.

In addition to using a five-point scale rather than a pass/fail scale, the TPAI is further distinguished from the CCI in that it is very prescriptive instructional methodology. For instance, for the indicator regarding the initiation of lessons (i.e., creating a structure for learning), the CCI relies on the observer's professional judgment regarding the appropriateness of the initiation in relation to the lesson objective(s). In contrast, for a similar indicator, the Georgia instruments specify four techniques to stimulate the interest of students, and then establish a rating procedure based on how many of the techniques are used; the more techniques observed, the higher the rating. This prescriptive definition of instruction does not allow for as wide a variety of teaching styles as does the CCI system.

The CCI, which has undergone extensive development and revision since 1985-86, is currently part of Connecticut's induction program for beginning teachers, the Beginning Educator Support and Training (BEST) Program. Connecticut is using the CCI to assess eligibility for provisional certification starting in 1989-90.

Although Connecticut requires that a beginning teacher be observed on six occasions by six different assessors, for California's pilot test the CCI was used for a single observation of each teacher. A single observation would not yield a sufficient sample of teaching evidence to make credentialing judgments. However, the focus of the California pilot test was to evaluate the instrument in more varied contexts than are available in the state of Connecticut; for this purpose, a single observation was judged acceptable. Also, a single observation per teacher was deemed sufficient for the purpose of trying out a high-inference classroom observation instrument, since much is already known about more behavioristic approaches.

The administration of the CCI in this pilot test, the content of the instrument, and the assessment format are discussed below. The content and format sections of the report contain information from the teacher and assessor evaluation forms, as well as information and analysis of scoring results. Following these three sections are sections on cost analysis and technical quality. The chapter concludes with an overall summary together with recommendations for further steps in exploring the feasibility and utility of high-inference classroom observation instruments such as the CCI in California teacher assessment.

Administration of Assessment

Beginning with an overview of the administration of the CCI, this section contains information on the following: logistics (e.g., identifying the teacher sample, scheduling classroom observations, etc.), security, assessors and their training, scoring, and perceptions of the instrument by teachers, assessors and FWL and RMC staff members.

Overview

The administration of an observation system, such as the CCI, in a new teacher's classroom requires careful planning on the part of the state, the observer, the new teacher, and the school administrator. Despite a very tight timeline, the use of trained assessors from the state of Connecticut made it feasible for FWL to complete the pilot test of the CCI during two weeks in May, 1989. As shown in Tables 3.1 and 3.2, the pilot testing was done in six California New Teacher Project locations, by six different trained assessors (who were at times accompanied by untrained independent observers from FWL and RMC), at different grade levels, and in several subject areas. Forty-one teachers participated in this pilot test.

Logistics

Administration of the CCI required the following logistical activities: identifying teacher samples, scheduling the observations, arranging for facilities in which to conduct the pre- and post-observation meetings, making travel arrangements for the Connecticut assessors (hereafter referred to as CT assessors), sending the orientation and CCI materials to the teachers, and acquiring evaluation feedback from the teachers and assessors.

As Table 3.1 indicates, the teacher sample for this assessment was almost equally divided between Southern and Northern California. Although we strove to ensure a roughly equal mix of elementary and secondary teachers, and a variety of teaching contexts, the highest priority became securing groups in compact geographic areas so as to reduce assessor travel time. In the Chico Project, for example, some of the rural schools participating in the Project are two to three hours apart.

After identifying the participants, teachers and principals were contacted to schedule the 45-60 minute observation and to arrange for a 15-20 minute pre-observation meeting and a 5-10 minute post-observation meeting at the school site. In accordance with the request of the experienced CT assessors, no more than two observations were scheduled for any one day, and two observations in one day were never scheduled two days in a row. (FWL and RMC staff who functioned as untrained observers for this assessment quickly discovered the necessity of this scheduling arrangement because scripting two lessons and completing two analytical write-ups in a single day was physically and mentally exhausting.) Although the CT assessors varied in their range of experience with regard to grade level and subject matter, it was not possible due to the small number of assessors to arrange for assessor-teacher matches along these dimensions.

Shortly before the observations, the participating teachers received orientation materials and a full copy of the CCI (including copies of the pre- and post-interview questions). Soon after the observations, teachers received an evaluation form to fill out and return to FWL. (Unfortunately, due to a clerical error, many of the teachers received the forms a month after the assessment. As a result, the return rate for the forms was low. Only 17 of the 41 teachers completed the forms.) Evaluation forms were also given to each of the CT assessors who returned them to FWL along with their assessment records.

Security

Because each teacher received a full copy of the CCI, the main focus of the security effort in this pilot test was on the completed documentation for each teacher. Assessors mailed the documentation materials to FWL, where they were securely filed.

TABLE 3.1

PILOT TEST PARTICIPANTS: CONNECTICUT COMPETENCY INSTRUMENT (CCI)

(Total Number of Teachers=41)

Dates	Project	Assessor	Number of Teachers	Teacher Characteristics
May 1–5	Long Beach	Bilingual Chapter I Teacher	7	3 Elementary; 4 Junior High 1 Male; 6 Female
May 1–5	Santa Barbara/ Ventura	Trainer of Trainers	7	7 Elementary 1 Male; 6 Female
May 8–12	Chico	Assistant Principal	7	5 Elementary; 2 High School 2 Male; 5 Female
May 8–12	Lodi	Instructional Consultant	7	7 Elementary 0 Male; 7 Female
May 8–12	Riverside/San Bernardino	Department of Education Staff	6	4 High School; 2 Middle School 1 Male; 5 Female
May 8–12	Winters	Higher Education Representative	7	2 Elementary; 3 High School; 2 Middle School 4 Male; 3 Female

TABLE 3.2

CONNECTICUT COMPETENCY INSTRUMENT: SUBJECTS BY GRADE LEVELS OBSERVED

SUBJECT	GRADES				TOTAL
	K-3	4-6	7-8	9-12	
Reading	2	4	1	1	8
Language Arts/English/Spelling	5	3*	2	2	12
Science	1	4	1	–	6
Social Studies	1	–	1	1	3
Mathematics	4*	1*	2	2	9
English as a Second Language	–	–	–	1	1
Music	1*	–	–	–	1
Health and Physical Education	1	1	–	–	2
Other Subjects	–	–	–	2	2
	=====	=====	=====	=====	=====
*Some observed lessons included multiple subjects.	15	13	7	9	44

If an observation system like the CCI is selected as a method of assessment for credentialing teachers in California, procedures to ensure security at the observation and processing stages (and during longer-term storage) would have to be developed and implemented by California. Each piece of documentation would have to contain identifying information (e.g., teacher code, observer code, date of observation) in case the pieces became separated. All documentation for a given teacher credential candidate would also have to be retained for a minimum number of years, enough to cover the period in which teachers could appeal decisions, or to meet statutory requirements.

Assessors and Their Training

Six trained assessors from Connecticut were invited to participate in this pilot test. The use of trained assessors from Connecticut rather than California assessors reduced the costs of the pilot test considerably. Had we recruited California assessors, it would have been necessary to train them. The use of trained Connecticut assessors also reduced the amount of staff time required to coordinate the pilot test (e.g., no recruitment or training was necessary), and enabled us to complete the pilot test in a relatively short period of time (two weeks).

In addition to already being trained, the Connecticut assessors had previous experience conducting the CCI assessment in Connecticut. This experience ranged from two assessors who had conducted three assessments of beginning teachers to one assessor who had conducted approximately twelve assessments.

Because the CCI design is based on the philosophy that the professional judgment of trained assessors is critical in making decisions about teacher competence, the CCI training process for assessors is an important component of the CCI system. While acknowledging that classroom teaching experience is a valuable basis for professional judgment, the creators of the CCI also realized that experienced educators have their own ideas and methods for determining effective teaching. The goal of the CCI system is to complement experience with training so as to ensure that the assessment criteria are consistently and objectively applied in rating teacher performances.

The training process for CCI assessors consists of five intensive days of instruction and practice, an independent field assignment, two days of follow-up instruction, and a proficiency test. During the five days of training, assessor candidates meet in groups of ten and work with two trainers to learn the following: the content and meaning of the ten indicators, the CCI standards, the procedures for conducting an observation, the skills necessary to document (or script) relevant information during the observation, and the skills necessary to write, weigh, and rate evidence from the scripted documentation. The training is conducted via whole- and small-group discussions and activities, with a focus on extensive daily practice in scripting and analyzing videotaped lesson segments representing one or more of the indicators.

Following the five days of training, the assessor candidates are given an independent field assignment which requires them to select and observe a teacher (someone not in Connecticut's beginning teacher program) and then to write and analyze evidence based on the observation. The results of this assignment are shared and discussed during the two days of follow-up training. Then the assessor candidates are given a proficiency test in which they analyze two videotapes of classes taught by beginning teachers.

A staff member from FWL and one from RMC participated in the five-day training sessions in the summer. Both had experience with the CCI when they functioned as independent observers in the spring pilot testing, so neither entered the training as a complete novice. Nevertheless, both found the intensive training to be stimulating and valuable. The daily discussions among participants, which centered on the content and meaning of each of the ten indicators of good teaching, as well as on how to write, weigh, and rate evidence for the indicators, were invaluable as a means of helping participants fully understand the meaning of the indicators and the rating standards. The continual professional interchanges also provided participants with fresh and stimulating insights into teaching in general and into their own teaching in particular. Both during and after the training, most participants claimed that, because of the training, they would be much better teachers when they returned to the classroom. Participants also expressed a renewed professional commitment to teaching and strong enthusiasm for participating in the process of inducting new teachers into the profession.

Although both FWL and RMC staff members found the training to be valuable, they also felt the training could be improved in at least two ways. These two factors could be instrumental in the event that California decides to develop a comparable system for its teacher certification process. First, more specific examples of written evidence for each indicator could be provided and utilized as part of the training. This would eliminate a lot of time spent by the trainees writing evidence inappropriately. (Written examples are available in the trainee's handbook, but these were seldom referred to by the trainers.) Second, the amount of scripting from videotapes could be reduced so that more time could be given to writing, weighing, and/or rating evidence. Although the daily scripting from videotapes shown on a medium-sized television screen provided beneficial practice in scripting, it was also an artificial situation. Scripting from a TV screen is not the same as scripting in a classroom. It is harder to "observe" the whole picture (i.e., the classroom and its participants) when the picture is limited to an area the size of the television screen. In addition, focusing on and scripting from a relatively small TV screen (compared to the size of the classroom) is very hard on the eyes. Instead of scripting as much, trainees could be given typewritten scripts for practice in writing, weighing, and/or rating evidence. In the latter two areas especially, some trainees experienced confusion even at the end of the training. Since the evidence component is one of the key features of the CCI that distinguishes it from other classroom observations systems, it is important that there is consistency among assessors.

Both staff members also considered the issue of shortening the training. The RMC staff member believes the training in Connecticut could be shortened if it is better organized (e.g., more specific examples are provided before exercises, more materials to read and study before the training). The FWL staff member agrees that the training would be improved with better organization, but is not sure the training should be shortened. As described above, in five days the assessor candidates are introduced to and expected to learn not only the content of the assessment and how to conduct the assessment, but also how to score the assessment. The participants are trained in how to be competent assessors and competent scorers at the same time. A total of five days training, which is approximately 2 1/2 days each for assessor training and scorer training covering 10 conceptually distinct indicators, does not seem to be an excessive amount of time. Moreover, because the CCI indicators must be applied to a broad range of teaching contexts, teaching styles, and pedagogical techniques, extensive discussion of applications of the indicators are necessary in order to ensure that observers are able to implement the CCI (or any high inference observation instrument) fairly. Finally, any training of California assessors in the use of a high inference observation instrument such as the CCI would also have to address the complexities of California classrooms (e.g., the diversity of students, large class sizes, the use of instructional aides). The length and structure of assessor training should be based on a careful evaluation of what skills are need by assessors to achieve a high degree of quality, consistency and reliability in the assessments.

In Connecticut, three types of assessors are used to administer the CCI: (1) state assessors, (2) administrator assessors, and (3) teacher assessors. Each beginning teacher is observed by two of each type for a total of six observations. The teachers participating in this pilot test were asked for their suggestions as to who should administer a classroom observation assessment (district administrators and assessors outside the district were given as examples of possible answers). The teachers' answers were as follows:

Assessors outside the district	9
Other teacher(s)	1
Site administrator	1
Other	3
No answer	3

Of the nine teachers opting for assessors outside the district, almost all did so because they believe such persons to be "less threatening" or "less intimidating," or because they would be "fair and unbiased," and there would be less chance of "playing favorite." The "Other" category included teachers who suggested that the instrument be administered by people who were well-trained in using it.

Scoring

The scoring system of the CCI is an integral part of the CCI process. That is, the same person who conducts the classroom observation uses the documentation from the observation to score the observation. The scoring system begins with a documentation form for each observation. The form requires the assessor to provide, from the scripted lesson, summaries of both positive and negative evidence for each of the instrument's ten indicators and their corresponding attributes. Each indicator has a page ("t-sheet") for recording the evidence. At the bottom of the page, the assessor is asked to consider the evidence in order to rate the teacher as "Acceptable" or "Unacceptable" on that indicator. (Some indicators also allow other ratings such as "Cannot Rate" or "Not Applicable.")

For this pilot test, two documentation forms were used: one an early version that was used in Connecticut pilot tests, and the other a revised version that had never been used before. The revised version differs from the older version in that it first asks the assessor to consider the evidence for each attribute and to give a rating of "Acceptable" or "Unacceptable" to that evidence. These attribute ratings are then combined, following rules established for each indicator, to obtain a rating of "Acceptable" or "Unacceptable" for the indicator. For example, the indicator, "Questioning Techniques," can only be rated "Acceptable" if all four of its attributes are also rated "Acceptable." The older version does not require a rating of the attributes, but only a consideration of the overall evidence corresponding to the attributes. According to Connecticut, the revised version makes it easier to rate the indicators because the decision rules are more clearly defined.

Upon completion of the evidence summaries and the individual indicator ratings, the assessor fills out a "Summary of Ratings" form which lists all the indicators (and on the revised version, the attributes) and shows the assessor's rating for each one. By looking at this "Summary of Ratings" form, one can determine how many "Acceptable" and "Unacceptable" ratings the teacher obtained. For certification purposes, the State of Connecticut requires teachers to obtain an "Acceptable" rating on at least seven of the ten indicators. (Certification in Connecticut, however, is *not* based on a single observation. A new teacher in Connecticut is observed six times: twice early in the year, twice in the middle, and twice near the end of the year. The two observations per time period are conducted by different observers, whose ratings are sent to an independent testing service. The testing service aggregates the ratings to obtain a single rating for each indicator. This aggregated set of ratings is sent to the teacher, who is urged, but not required, to share the results with the assigned mentor teacher.)

The CCI scoring system is very labor intensive. The entire process takes from two and one-half to four hours per assessment because the assessor/scorer must write up at least ten pages of evidence (one for each indicator) and then carefully analyze the evidence in order to give a rating to each of the attributes and the indicators.

Teacher, Assessor, FWL, and RMC Staff Perceptions of Administration

As reported in the Spring 1989 Administration Report, the majority of teachers were satisfied with the administration of this assessment. Half of the assessors were also satisfied, while half did not like the number of observations scheduled for the week. Said one assessor:

> *Seven assessments in one week is unrealistic. The quality of an assessor's write-up is directly affected as the number of observations increases beyond three a week. However, if an assessor is observing as their only occupation, one a day is feasible.*

FWL and RMC staff who served as independent observers for this assessment concurred with the above observation, and also noted that if classroom observations were selected as a method of assessment for credentialing teachers in California, assessor fatigue resulting from traveling between school sites, especially in rural areas, should also be considered.

The assessors reported two difficulties in administering this assessment: (1) the amount of time it takes to write evidence and rate an observation, and (2) not being able to give teachers some feedback after the observation. Several of the teachers also expressed a strong desire for feedback.

FWL and RMC staff also found the amount of time to write evidence (i.e., up to four hours) to be a difficult part of the administration of this assessment. In addition, as mentioned earlier, the logistics of scheduling the teachers in rural areas presented some difficulty. Many rural schools are so far apart that even scheduling one observation a day required careful calculation.

Assessment Content

The content of the CCI focuses on teaching behaviors that are directly observable in the classroom: management of the classroom environment, instruction, and assessment of student progress. The content is firmly grounded in the research literature on effective teaching, and it incorporates the experience and ideas of Connecticut teachers, district administrators and teacher educators.

The development of the CCI content stemmed from a 1984 validation of the Connecticut Teaching Competencies (Streifer 1984). In 1985-86, a grant from the National Institute of Education to evaluate the feasibility of establishing an induction program for Connecticut's beginning teachers led to a first draft of an assessment instrument. This instrument was greatly modified after a 1987 conference in which national experts met with Connecticut State Department of Education (CSDE) staff to discuss philosophy and approaches to performance assessment, instrument development and standard setting. As a result of this conference, a small working group of practitioners, CSDE staff and researchers created a first draft of the CCI.

A second panel of national experts in teacher assessment, observation methodology, research design, and implementation of state assessment programs critiqued the draft CCI in 1987. The instrument was given a small-scale pilot test in December of that year. At about the same time, the draft was critiqued by Connecticut representatives of higher education, professional organizations, local district staff and state department personnel. More revisions were made to the CCI, another small-scale pilot test was conducted in 1988, and, after more revisions, a full pilot test was conducted with 220 beginning Connecticut teachers in 1988-89.

Also in 1988, over 1,500 Connecticut educators participated in a content validity study in which they rated the appropriateness of the CCI's indicators to the job of teaching in Connecticut. As part of this validity study, the generalizability of the instrument was also evaluated, and a bias review was completed.

The content of the CCI was still being revised in 1989. As mentioned earlier, two versions of the CCI were used for this pilot test: (1) a version that does not include ratings by attributes and that had been used in previous Connecticut assessments, and (2) a recently revised version that does include ratings by attributes and that had never been used. Because only two of the CT assessors used the older version, this section focuses on the more recent version. Although the two forms differ in some of the attributes which define each indicator and in the criteria for rating, both forms focus on the same ten indicators named above.

In the following pages, the content of the CCI is evaluated on the basis of seven factors:

- o Congruence with California curriculum guides and frameworks;
- o Extent of coverage of California Standards for Beginning Teachers;
- o Job-relatedness of the instrument;
- o Appropriateness for beginning teachers;
- o Appropriateness across different teaching contexts (e.g., grade levels, subject areas);
- o Fairness across groups of teachers (e.g., ethnic groups, gender); and
- o Appropriateness as a method of assessment.

We would like to note that, just as Connecticut educators reviewed the CCI for job relevance and importance, if the CCI is to be further field tested in California, a validity study should be done at the same time. (For more on validity, see the section, Technical Quality.) Without such a study, and with our pilot test sample of 41 teachers, our ability to comment on the CCI's appropriateness along such dimensions as job-relatedness, appropriateness for beginning teachers, and appropriateness across contexts is limited. Thus, except for the first two dimensions of curriculum congruence and standards coverage, the discussions of the remaining dimensions are based on the perspective of the participating teachers, the CT assessors, and FWL and RMC staff, as reflected in feedback forms, in informal conversations with the assessors, in meetings with and a report from RMC staff, and in data from the CCI ratings sheets.

Congruence with California Curriculum Guides and Frameworks

The California curriculum guides and frameworks are, by definition, subject specific, which the CCI is not (the CCI focuses on generic teaching behaviors which can be applied across subjects). Nevertheless, FWL staff looked at the CCI to see if there is congruence with the guides and frameworks, and how the CCI could be modified to improve congruence. For our analysis, we examined the following four California guides and frameworks: *English-Language Arts Guide*, *Mathematics Framework*, *Science Guide*, and the *History-Social Sciences Framework*. Because the guides and frameworks were developed independently by subject-matter panels, they vary markedly in their foci and degree of specificity. We did not look at the curriculum guides in other areas, but we would expect similar results to those discussed below.

Table 3.3 briefly describes the content of each of the four guides and frameworks, and also lists the CCI indicators which address the content. As the table indicates, there is only partial congruence between the CCI and the guides and frameworks. It should be noted, however, that the CCI is a generic, non-curriculum specific, high - inference observation system. As such, it does not measure a teacher's knowledge of curriculum directly. On the other hand, the CCI includes several indicators to assess a teacher's presentation of the content of a lesson: "Structure for Learning," "Lesson Content,"

TABLE 3.3

CONGRUENCE OF CCI WITH CALIFORNIA CURRICULUM GUIDES AND FRAMEWORKS

Curriculum Guide or Framework	Content Description	Relevant CCI Indicators	Comments
1. English-Language Arts Guide	22 guides for instruction in grades K-8.	None	None
2. Mathematics Framework	5 major emphases of curricular content.	None	None
	10 characteristics of instruction.	Lesson Content, Questioning Techniques, Monitoring, and Adjusting	The three indicators are congruent with the three characteristics: Mathematical Language, Questioning and Responding, and Corrective Instruction.
3. Science Guide	Content-knowledge descriptions of biology, earth, and physical science programs for grades K-8 (includes ideas on how to teach the subject matter).	None	
	General characteristics of a strong science program.	Lesson Content, Questioning Techniques	Indicators corresponding to characteristics focus on development of students' emotional, physical, and intellectual development and questioning techniques and responses.
4. History-Social Sciences Framework	Three curricular goals, their corresponding learning strands, and a sequential curriculum for grades K-12.	None	None

"Lesson Development," "Questioning Techniques," and "Communicating Clearly." Assessment of the content through these indicators could be strengthened if the observation were conducted by an observer with special expertise in the subject matter of the lesson. To obtain more extensive evidence of a teacher's knowledge and ability to teach specific curriculum content, other measures would be needed. These measures could be alternatives to observations, such as interviews, written assessments or combinations of these with an observation system.

Extent of Coverage of California Standards for Beginning Teachers

As mentioned earlier, the CCI was specifically developed with the Connecticut Teaching Competencies in mind. It was not developed to assess the standards for beginning teachers that have been established by California's Commission on Teacher Credentialing. Although there are similarities between the Connecticut competencies and the California Standards, they are not identical. FWL staff examined the CCI indicators to see how well they assess Standards 22 through 32 of the California Beginning Teacher Standards, which define levels of pedagogical competence and performance that California teacher credential candidates are expected to attain. These standards are reprinted below (in italics), along with an analysis of how the CCI indicators correspond to each standard.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* The CCI indicator, "Positive Learning Environment," requires the observer to look for evidence of student rapport and of a classroom environment that is conducive to learning. The indicator, "Behavior Standards," requires evidence that the teacher has established standards of student behavior and applies fitting consequences to both appropriate and inappropriate behaviors.

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies, activities, materials and assessment plans that are well defined and coordinated with each other.* The CCI process requires the teacher to plan a 45-60 minute lesson for observation and to specify on a pre-observation form the objectives, activities, instructional arrangements, and materials that are part of the lesson. In addition, the indicator, "Lesson Development," requires the observer to look for evidence that the teacher has developed the lesson in a logical or sensible order, and that the materials and instructional arrangements used for the lesson are consistent with the planned or emerging lesson.

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities and materials that are appropriate for students with diverse needs, interests and learning styles.* The CCI indicator, "Lesson Content," requires the observer to seek evidence that the teacher's choice of content (defined by the instrument as "student learning activities, lesson materials, teacher presentation, and teacher questioning" as manifested in the lesson) is appropriate to the students' level of development. This indicator ***does not*** directly address, however, the issue of students with diverse learning styles and interests, although evidence for this standard may be found during the observation. It also does not assess whether the teacher's strategies, techniques, and materials are "free from bias."

Standard 25: Student Motivation, Involvement, and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* Several CCI indicators address this standard: The indicator, "Positive Learning Environment," requires the observer to look for evidence that the teacher "creates a climate that encourages all students to achieve"; the indicator, "Appropriate Standards of Behavior," asks the observer to look for evidence that the teacher "communicates and reinforces appropriate standards of behavior for the students"; the indicator, "Student Engagement," requires the observer to look for evidence that the teacher involves "a clear majority (at least 80%) of the students in the instructional activities of the lesson"; and the indicator, "Appropriate Questioning Techniques," asks the observer to seek evidence that the teacher, through questioning techniques, provides opportunities for most students (including those of different ethnic groups and genders) to be involved in the lesson.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* The CCI indicator, "Communication Skills," requires the observer to seek evidence that the teacher communicates in a "coherent manner, avoiding vagueness and ambiguity that interfere with student understanding." This indicator also assesses the teacher's technical quality of communication, focusing on articulation, volume, and rate of delivery. This indicator does not address, however, the teacher's written language, and so the indicator would have to be changed to include a focus on the teacher's written language in order to meet the standard. This standard is also addressed by the CCI indicator, "Appropriate Lesson Content," which asks the observer to ascertain if the teacher uses "vocabulary and language appropriate to the learners."

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* The CCI indicator, "Positive Learning Environment," asks the observer to supply evidence that the teacher "creates a climate that encourages all students to achieve" (i.e., communicates expectations for achievement). The indicator, "Monitoring and Adjusting," asks the observer to look for evidence that the teacher "checks the level of student understanding at appropriate points during the lesson," and, when monitoring indicates that students are misunderstanding or failing to learn, or that students have mastered the concepts being taught, that the teacher uses "appropriate strategies to adjust his or her teaching." The CCI does not assess the methods a teacher uses to ascertain students' prior attainments related to the subject of the lesson or the methods used to formally evaluate student work.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* The CCI indicator, "Structure for Learning," requires the observer to find evidence that the teacher's lesson includes closure(s) which could help the students to evaluate information, think analytically, and reach sound conclusions. It does not evaluate a teacher's ability to design instruction that increases the critical thinking skills and problem-solving ability of students unless that is the objective of the lesson observed; if that *is* the lesson's objective, then the indicator, "Questioning Skills," requires the observer to find evidence that the teacher asks high-order cognitive questions.

Standard 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward the subjects learned, the students themselves, and their capacity to become independent learners.* The CCI indicator, "Positive Learning Environment," requires the observer to find evidence that the teacher demonstrates "patience, acceptance, empathy and interest in students through positive verbal and non-verbal exchanges;" "avoids sarcasm, disparaging remarks, sexist or racial comments, scapegoating or physical abuses;" "exhibits her or his ***own*** enthusiasm for the content and for learning;" and "maintains a positive social and emotional tone in the learning environment." It does not assess, however, whether a teacher encourages positive interaction ***among students*** or independent learning experiences.

Standard 30: Capacity to Teach Cross-culturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural, gender, linguistic and socio-economic differences.* Although no CCI indicator addresses this standard directly, the CCI process and several indicators (e.g., "Positive Learning Environment," "Lesson Content," and "Questioning Techniques") allow the observer to note whether the teacher demonstrates rapport with, and the ability to teach, students who are different from the teacher. If, however, the classroom is homogeneous with respect to ethnicity or culture or socioeconomic differences (e.g., several classrooms in Northern California appeared to be ethnically homogeneous), then the CCI cannot even indirectly assess this ability.

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* This standard focuses on a teacher's ability to teach classes which span the range covered by the credential (i.e., grades K-8 or 7-12) or students at two or more ability levels (such as remedial and college preparatory classes). None of the CCI indicators are designed to assess this ability. (It would be possible, however, to compare the observations of a teacher who teaches both remedial and college preparatory classes.) This standard also addresses a teacher's ability to fulfill typical responsibilities of teachers such as meeting school deadlines and keeping student records, none of which are assessed by any CCI indicator.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and develops professionally through self-assessment and collegial interactions with other members of the profession.* None of the CCI indicators assess whether a teacher fulfills his/her obligations as a member of a profession and a school community (e.g., adheres to high standards of professional conduct and engages in collegial relationships).

The extent of coverage by the CCI of the California Beginning Teacher Standards is summarized in Table 3.4. The table lists the CCI indicators that address each standard and also describes the extent of coverage provided.

TABLE 3.4

EXTENT OF COVERAGE BY THE CCI OF CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Standard	CCI Indicator(s) Assessing Standard	Extent of Coverage
22: Student Rapport and Classroom Environment	–Positive Learning Environment –Behavior Standards	Full
23: Curricular and Instructional Planning Skills	–Lesson Development	Partial
24: Diverse and Appropriate Teaching	–Lesson Content	Partial
25: Student Motivation, Involvement, and Conduct	–Positive Learning Environment –Behavior Standards –Student Engagement –Questioning Techniques	Full
26: Presentation Skills	–Lesson Content –Communication Skills	Partial
27: Student Diagnosis, Achievement, and Evaluation	–Positive Learning Environment –Monitoring and Adjusting	Partial
28: Cognitive Outcomes of Teaching	–Structure for Learning –Questioning Techniques	Partial
29: Affective Outcomes of Teaching	–Positive Learning Environment	Partial
30: Capacity to Teach Crossculturally	–None	None
31: Readiness for Diverse Responsibilities	–None	None
32: Professional Obligations	–None	None

Job-relatedness

Sixteen of the 18 teachers who evaluated the CCI stated that all the major competencies measured by this assessment are relevant to their job of teaching. Some of the teachers described the job-relatedness of the CCI as "excellent." Other teacher comments were as follows:

> *All areas are relevant to my job of teaching.*
>
> *I believe that the instrument covered all areas.*
>
> *Yes! I felt that my* **entire** *mode of teaching was being evaluated, not just my lesson.*

Because the CCI assessment entails observing teachers actually teaching in their own classrooms, the job-relatedness of this assessment is strong. Job relevance is a particularly important factor in evaluating different approaches to teacher competence assessment, because professional practitioners and courts of law consider this factor first when they judge the fairness of an evaluation system. Furthermore, as a classroom observation system, the CCI offers direct evidence of actual teaching competence. For this reason, it is not necessary to make inferences about how well a teacher conducts instruction if such an assessment is used. Making inferences about the quality of a teacher's actual teaching is a primary characteristic of all other approaches to teacher assessment, as will be shown in subsequent chapters of this report.

Appropriateness for Beginning Teachers

Teachers were asked if they felt they had an opportunity to acquire the knowledge and abilities measured by the CCI. A slight majority of teachers (10 of 18) responded affirmatively. One teacher remarked:

> *Experience will certainly help a teacher become more effective in all areas but this is good for the beginning teacher to begin to focus on the specific skills listed in the assessment instrument.*

Three teachers responded negatively to the appropriateness for beginning teachers question; five teachers either did not respond or gave answers which did not address the question.

The CT assessors were also asked if they thought the CCI assessment was appropriate for beginning teachers. Their responses were generally affirmative -- but with qualifications. Several of the assessors stated that the assumption that beginning teachers can acquire the knowledge and skills needed to demonstrate competence by the end of the first year is a valid one, but only if the teachers have received mentoring, supervision and support during the year as they do in Connecticut. Explained one assessor:

> *Although the indicators are written with vocabulary and terminology that is global, the variety of ways each indicator*

> *can be expressed in terms of specific behaviors is best addressed cooperatively so that someone with more extensive classroom experience can broaden a beginning teacher's experience.*

Some assessors also mentioned that the California teachers had particular difficulty with the indicators "Structure for Learning" and "Questioning," and, as a result, questioned whether the teachers received sufficient training in these areas. The assessors perceived the teachers' difficulty with these indicators as further indication that teachers need more assistance or preservice education.

Our analysis of the CCI ratings (scores) revealed that, for the most part, the beginning teachers performed well on the CCI. Approximately 80% (33 of the 41 teachers) received "acceptable" ratings on at least seven of the ten indicators. Almost 40% (16 teachers) received "acceptable" ratings on all ten indicators. Of the teachers who did not perform as well, approximately 15 percent (six teachers) received four to seven "unacceptable" ratings, and about five percent (two teachers) received as many as eight to nine "unacceptable" ratings. No teacher was rated as "unacceptable" on all ten indicators.

Our scoring analysis also suggested that of the ten indicators, teachers had the most trouble with one indicator in particular. As Chart 3.2 indicates, of the 41 teachers who were observed, only 20 teachers (49%) received an "acceptable" rating on the "Structure for Learning" indicator. This may suggest that many beginning teachers need more training or experience in providing initiations and closures to lessons (i.e., "Structure for Learning). Alternatively, this skill may develop with experience, so the CCI standards for this teaching ability (indicator) may be inappropriate or too high for **beginning** teachers. (It is the opinion of FWL staff, however, that the first explanation is more likely than the second.)

Appropriateness across Contexts

Almost all of the teachers (15 of 18) felt the CCI approach is useful for teachers in different contexts (i.e., across grade levels, subject areas, and diverse student groups). Some of the teachers, however, qualified their answers, saying that its usefulness was dependent upon the teacher receiving feedback after the observation; for example, one teacher remarked, "It could be [useful] if I actually got feedback from it." The following pages examine the issue of the CCI's appropriateness for different grades, subjects, and student groups.

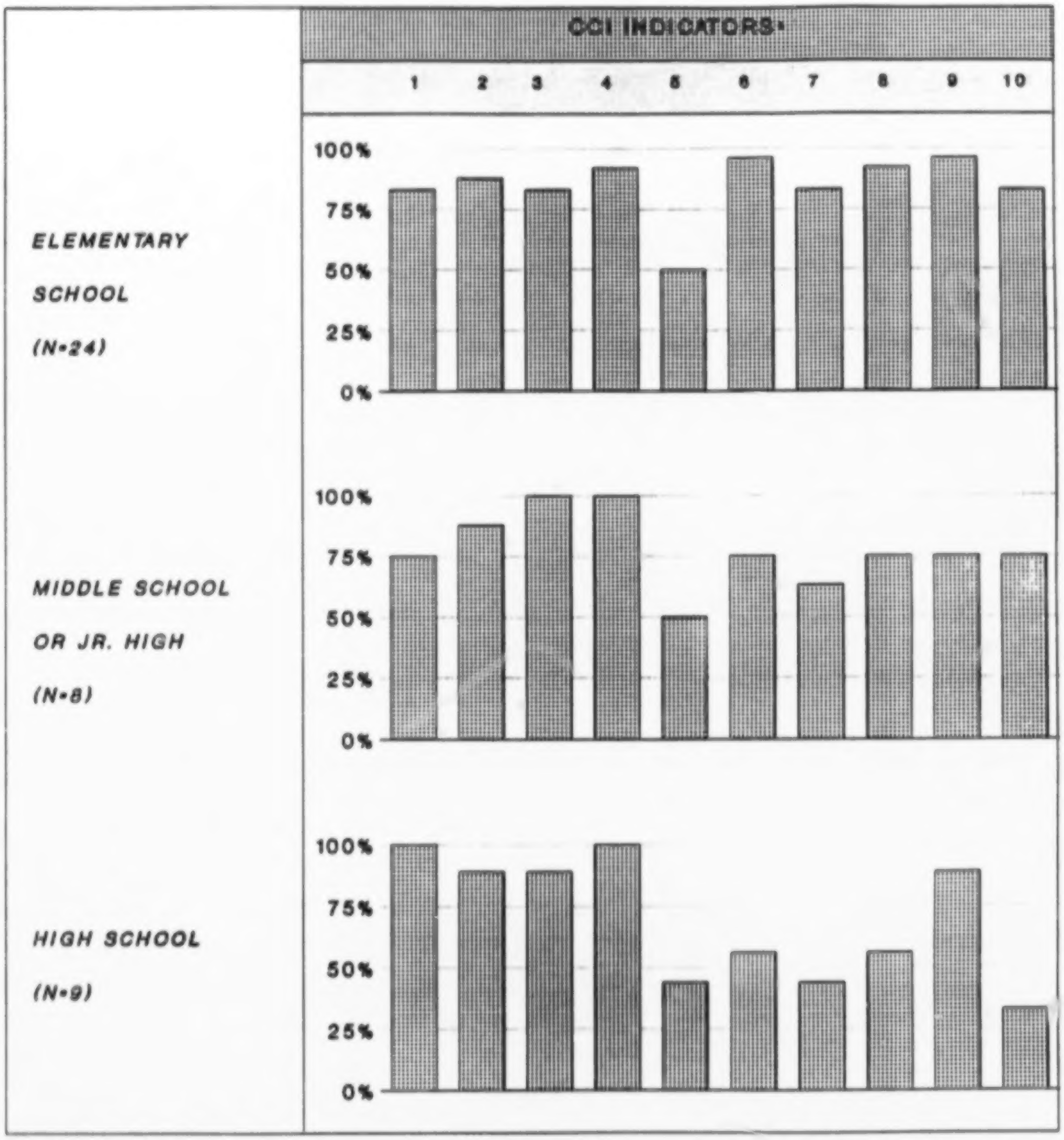
Grade level and subject matter. As indicated above, most teachers believe the CCI is appropriate for new teachers in varying grades and subject areas. (See Table 3.2 for a complete listing by grade level of all subjects observed.) This opinion was also expressed by almost all of the CT assessors. Commented one assessor:

> *Considering that I have teaching experience spanning preschool through 11th grade, I am confident that this instrument is relevant to* **all** *grades and subject areas.*

CHART 3.2

CCI SCORING RESULTS

PERCENT OF TEACHERS RECEIVING "ACCEPTABLE" RATINGS PER INDICATOR



* 1=Learning Environment
2=Behavior Standards
3=Student Engagement
4=Routines/Transitions
5=Structure for Learning
6=Lesson Content
7=Lesson Development
8=Questioning
9=Communication
10=Monitoring and Adjusting

One assessor, however, questioned whether the instrument was appropriate for all subject areas. This assessor observed a teacher in an agricultural management class and found that "no teaching of the sort defined by the Instruction and Assessment clusters occurred." According to her report, the observed lesson did not contain any explicit instruction from the teacher. Instead, the teacher acted as a supervisor while the students engaged in a particular activity (in this case, worming sheep).

The assessor stated that based on her many years of experience as a high school administrator observing industrial arts and business classes, many -- if not most -- vocational classes are like this; that is, the approach seems to be, "Just do the task and you'll learn." Exactly what is learned, however, is not necessarily specified. Although the observed lesson took place on what the teacher described as an "activity day," the assessor asked if there aren't some classes (e.g., vocational education classes) that have less emphasis on direct instruction than do other classes. She questioned whether these classes should be assessed in the same manner as other classes. In other words, is it appropriate to rate the teacher of the agricultural management class on the use of 1 questioning techniques, for example, when the nature of the lesson, as perceived by the teacher, required little questioning? Or is the definition of good teaching such that the teacher should have been expected to use questioning techniques during a "hands on" learning activity? Based on one's answers to these concerns, the CT assessor suggested that the CCI may or may not be appropriate across subject areas.

To begin to address the question raised by this assessor, we examined the documentation for the teacher who taught the agricultural management lesson. Although the assessor checked a "Cannot Rate" for lesson content, there were four other areas in which the teacher was given an "Unacceptable" rating: "Structure for Learning", "Lesson Development", "Questioning Techniques", and "Monitoring and Adjusting." The evidence given for the negative ratings seemed to justify the ratings. For example, the teacher's initiation and closure ("Structure for Learning") were purely administrative (e.g., after the last lamb was vaccinated, the teacher said, "That's it. Go in and get washed up."). Even a very activity-based lesson such as this (i.e., worming sheep) would benefit from a structural framework which facilitates learning. Another example is that although the teacher stood in the presence of the students and could see if they were performing the procedure correctly, he was not observed monitoring whether or not they understood the procedure. Such monitoring would also facilitate learning, even in an activity-based class. Thus, although the data are too limited to draw a definitive conclusion on this issue, the CCI does seem capable, at least to some degree, of assessing teachers of a variety of different subjects, including vocational education and other activity-based classes.

In addition to looking at the particular documentation for the lesson in question, we also looked at the distribution of ratings across subject areas of the teachers who received four or more "Unacceptable" ratings (i.e., failed the assessment by Connecticut standards). The teachers who fell into this category included four of the 19 teachers who were observed teaching English/language arts (includes reading), two of the six teaching science, one of the three teaching social studies, and the one teaching agricultural management. Moreover, of the four English/language arts lessons, two were taught by elementary teachers and two by high school teachers. Although the data are too limited to draw any conclusions about the relative degree of difficulty across subject areas, there was only one instance where an assessor reported difficulty in applying the CCI to the observed lesson, and that was for the agricultural management class.

In connection with this question of appropriateness across subjects, the CT assessors were asked how much knowledge of the subject matter should an assessor have to administer this assessment. Several of the assessors felt that there are certain content areas where a teacher-assessor match is a must. In particular, the assessors noted the importance of subject-matter knowledge at the secondary level. One assessor stated that her knowledge of physical science was crucial to her analysis of a middle school science lesson, and her lack of content knowledge was an impediment to her analysis of a high school vocational agriculture lesson. This assessor also questioned her ability to judge the level of difficulty of a first-grade lesson considering that her background was as a chemistry/physics teacher and high school administrator.

Two of the assessors, however, strongly dissented. One stated that knowledge of the subject matter is not essential to administering this assessment, but that knowledge of the instrument is. The other maintained that through focusing on cues such as the pattern of student engagement (i.e., are ***most*** students wearing puzzled expressions and withdrawing from active participation in the lesson?) and the content of questions asked by students, she could infer whether or not the lesson content and development were satisfactory. This particular assessor had experience teaching in both secondary and middle schools, so it is possible that she was speaking from a broader range of experience than other assessors.

Although there are merits to both sides of the argument, it would be difficult to judge the accuracy of the lesson content (as required by the indicator, "Appropriate Lesson Content") or the appropriateness of the questions asked to the lesson content (as required by the indicator, "Questioning Techniques") if one is not familiar with that content. Further, how could an observer judge the communication skills (as required by the communication indicator) of a teacher using a language other than English during much of the observation period if the observer does not speak that language? An observer does have the option of giving a "Cannot Rate" rating to the "Lesson Content" indicator, but not to the other indicators. Thus, it seems desirable that, whenever possible, the observer have some familiarity with the subject area in order to provide higher quality observations and more reasonable ratings on the attributes and indicators. In addition, if feedback is to be provided to the teacher, it would probably be more useful if it is based on the observation(s) made by someone familiar with the subject.

In our analysis of the CCI ratings, we also focused on the appropriateness of the CCI assessment across grade levels. We found that on some indicators there was a difference in how well teachers at different grade levels (i.e., elementary, middle, and senior high school) performed. Chart 2 shows the total number of teachers who received an "Acceptable" rating for each indicator, and also the number and percentage of teachers with the rating at the elementary, middle school, and high school levels. As the table shows, our sample of high school teachers tended to perform less well than our sample of middle school and elementary school teachers on four of the ten indicators. The percentage of middle school and elementary school teachers receiving an "Acceptable" rating was much higher than the percentage of high school teachers for the following indicators: "Lesson Content," "Lesson Development," "Questioning Techniques," and "Monitoring and Adjusting." In addition, one third of the high school teachers received an overall total of from four to seven "unacceptable" ratings compared to only one of the eight middle school teachers and two of the 24 elementary teachers. Due to the small size of each group, no conclusions can be drawn from the data. It may be beneficial, however, for any further pilot testing of high-inference observation instruments to include a focus on grade level comparisons. For example, future pilot tests could be conducted to see if a match between teacher and assessor with regard to grade level experience yields the same or different results as this pilot test. Similarly, additional pilot tests might be conducted to assess whether an assessor-teacher match with regard to subject matter affects results by grade level, particularly at the high school level.

Diverse students. The philosophy of the CCI, as stated in the CCI training handbook for assessors, includes the assumption that "Effective teaching is sensitive to cultural diversity." The handbook also states that "Competent beginning teachers will help prepare children for participation in a culturally diverse world and will also teach in ways that help all children learn." To the latter end, the CCI puts a special emphasis on the concept, "all children." For example, the effective teacher is expected to be accepting of and interested in all students, to encourage all children to achieve at the highest level they can, to engage all children in the learning activities, and so on. However, as noted in the earlier discussion of how well the CCI addresses "Standard 30: Capacity to Teach Cross-culturally" of the California Beginning Teacher Standards, the CCI does not completely capture a teacher's capacity to respond appropriately to diverse students because its sole focus is on those students present in the teacher's current classroom. If those students are all or mostly homogeneous, the CCI is completely unable to assess how a teacher responds to diverse students. Furthermore, although the underlying philosophy of the CCI may include the assumption that a competent beginning teacher will help prepare children for participation in a culturally diverse world, there is nothing in the content of the CCI that assesses a teacher's ability to do so.

The CCI could, however, be modified in order to address these issues. For instance, the first indicator, "Positive Learning Environment," might be modified to include a stipulation that, not only does the teacher "maintain a positive social and emotional tone in the learning environment," but also maintains a learning environment that is both fair to different types of students--by gender, ethnicity, handicapping conditions, language group, etc.--and is reflective of a culturally diverse world.

To further strengthen the CCI's capacity to assess a teacher's ability to teach diverse students, it also seems desirable to have, whenever possible, an observer who is familiar with the type of student group being observed. For example, special education students, limited English proficient students, and some students of particular ethnic groups may tend to exhibit certain characteristics. An observer's knowledge of student development and of desirable and appropriate behaviors for those in the classroom being observed would likely contribute to higher quality observations.

Fairness across Groups of Teachers

The majority of teachers responded positively to the question of fairness of the CCI across groups of teachers (e.g., different ethnic groups, different language groups, etc.). Thirteen of the 18 teachers believed the assessment is fair, two did not, and three did not give an answer (or gave an answer that did not address the question). The six Connecticut assessors also felt the CCI is fair across groups of teachers.

FWL staff are unable to comment on the fairness of the CCI across groups because there is not enough information about the teachers' ethnic backgrounds, language abilities, etc., to enable us to examine teacher performance with regard to these dimensions.

Areas of Most/Least Emphasis

Because the CCI assesses a variety of areas, the teachers were asked what areas they feel should receive the most/least emphasis in making decisions on credentialing. The teachers gave a wide variety of answers (there were also eight teachers who did not answer), some of which, together with the number of teachers who gave them, are listed below:

Most Emphasis

-- the way a teacher relates to students	(3)
-- teaching methods or instructional techniques	(3)
-- major competency areas	(1)
-- positive attitude	(1)
-- accuracy	(1)
-- flexibility	(1)
-- student engagement	(1)
-- the tone of voice in the classroom	(1)
-- monitoring for understanding	(1)

Least Emphasis

-- classroom management	(2)
-- style	(1)
-- high level knowledge of content	(1)

The majority of CT assessors generally felt that none of the areas should receive most/least emphasis. In defense of their opinion, two of the assessors made reference (either directly or indirectly) to the "integrated, holistic nature of teaching," and thus the importance of all the areas. One assessor explained, "A teacher with a dynamite plan but no discipline is no more effective than a teacher with great discipline and no plan."

One assessor disagreed and specified three areas she thought should receive the most emphasis: maintaining appropriate standards of behavior, promoting a positive learning environment, and monitoring student understanding and adjusting teaching. Regarding the latter, she commented, "No matter how good the teacher thinks the lesson is, without monitoring and adjusting she'll never know."

Assessment Format

Traditionally used by school administrators, the classroom observation method of assessment is generally accepted by teachers, administrators, parents, and the general public as an appropriate method to assess teacher competence. It is relatively easy to administer because it requires minimal materials (paper and pen for the assessor) and no special setting. Moreover, the person making the observation usually focuses on one specific area or takes general notes on a variety of areas. The classroom observation assessment as described by the CCI is also fairly easy to administer, but its format renders it more difficult than most traditional systems because the CCI requires the assessor to script, as best as possible, the entire lesson. In addition, the CCI analysis entails much more writing than traditional systems, and, perhaps, more careful codification.

From another perspective, the CCI cannot be easily administered to groups of teachers because its format requires one assessor observing one teacher at a time. Because the format requires the assessment to take place in the teacher's classroom (which is convenient for the teacher), the assessor must be able to travel whatever distance necessary to observe at the teacher's school site. Needless to say, in the state of California, these format issues pose a formidable challenge.

Other format issues more easily addressed are the clarity of the assessment, the clarity of the assessment materials, and the question of giving feedback as part of the assessment.

Clarity of the Assessment

Before the observations, each participating teacher received the Connecticut Competency Instrument which described the aspects of teaching being assessed (i.e., the 10 indicators). However, only nine of the 18 teachers who returned the evaluation forms responded positively when asked if they knew what aspects of teaching were being measured by the assessment (six teachers said "no," and three did not respond). Four of the nine teachers who said "yes" were also able to identify the aspects that they believed were being measured. Of the six teachers with negative responses, two expressed some confusion as to what was being evaluated -- the teacher or the CCI?

The responses to the above question lend credence to the recommendation made by some of the assessors that the teachers have assistance in reviewing the instrument. Eleven pages of description of the CCI indicators are probably too much for a teacher to understand through reading alone. It seems important that a teacher about to be assessed with the CCI have the opportunity to discuss the CCI process and the instrument itself with someone who is familiar with the CCI assessment.

Clarity of Assessment Materials

The majority of teachers (14 of 18) reported that the CCI assessment materials they received prior to the assessment were helpful. An even larger number of teachers (16) stated that the sample pre-assessment information form was helpful. Teachers were also asked to comment specifically on the pre- and post-observation forms. Almost all of the teachers (15) found the questions on the forms to be understandable. One teacher had praise for both the questions in general and the post-observation questions in particular:

> *The questions were clear; I liked being able to explain special circumstances or changes in plans.*

Based on the comments from the teachers and CT assessors, the pre-assessment information form and the pre- and post-observation interview forms seem to be especially valuable and key to the observation and evaluation of the teacher's behaviors. The pre-observation form gives structure and meaning to the teacher's lesson, and the post-observation form allows the assessor to understand the teacher's response to the classroom context, and to gauge the extent to which instruction was altered according to that context. If California decides to use a classroom observation instrument in teacher credentialing, there are several ways in which the materials should differ from the CCI materials.

Above all else, an identification number should be assigned to each new teacher to permit the linking of various documents and it should appear on each page of all the forms. Other suggestions for changes to some of the forms are as follows:

Pre-assessment Information Form -- Change so that it collects information on the name of the school, the name of the city, and the composition of the class (e.g., gender, ethnic groups, languages spoken, students receiving special services). Revise question #3 of this form, which asks what other adults will be in the classroom during the observation, to ask about other students that are not part of the teacher's regular class. Require, whenever possible, that the teacher attach relevant materials, such as copies of worksheets that are connected with the observed lesson.

Pre-observation Interview Form -- Rewrite for better clarity. Question #1, for example, asks the teacher if she/he has made any changes to the lesson plan described in the Pre-Assessment Information Form. If the teacher's response to that question is that there are no changes, Question #3 might be confusing because it asks the teacher about "other changes" that the interviewer should be aware of.

Scripting Sheets -- Current scripting sheets are notepad pages of the assessor's choice to which the assessor adds columns to match the CCI format. Replace with preprinted scripting sheets designed to conform to the official format. Include a space to collect information on the number of students on task, which would also serve to remind the observer to collect this information on a regular basis.

Assessor Rating Summary -- Modify to include a place to indicate if there is an Incident Report or not. (An Incident Report is completed by the observer if there is any irregularity that could affect the validity and accuracy of the observation and ratings.)

Documentation Checklist -- Connecticut provides a checklist for observers, listing their procedures and responsibilities. Prepare a similar document for California observers including timelines and addresses for sending documentation, a one-page list of clusters, indicators, and attributes in outline format with codes assigned to each, and a list of standard abbreviations for use in scripting to ensure some consistency across observers and for use in interpreting and reviewing another person's script.

Observation Feedback

The format of this assessment did not include giving feedback to the teachers, because the purpose of the pilot test was to evaluate the CCI instrument. Almost all of the teachers (16 of 18) indicated they would have liked some feedback, and two of the Connecticut assessors indicated that not being able to give the teachers feedback was the most difficult (or one of the most difficult) aspects of this pilot study.

The teachers were asked to describe what kind of feedback would be helpful from this type of assessment, and also by whom, when, and in what format the feedback should be provided. The most common response was that the assessor should provide the feedback (10 teachers), as soon as possible (9 teachers), and in a constructive, positive form (7 teachers).

A few teachers were not so concerned with how or what kind of feedback be given, just that it be given. Said one teacher:

> *Some feedback would be nice, in* **any form.**

Clearly, teachers desire feedback. However, the content of and process for providing observation feedback need to be carefully considered. In Connecticut, the feedback process for 1989-90 will consist of the beginning teacher receiving feedback after each observation. The feedback will provide information on whether the teacher demonstrated sufficient or insufficient skills relating to each defining attribute, or whether there was insufficient evidence to arrive at a conclusion. If California considers adopting a similar observation feedback process, feedback should be given relatively soon after the lesson so the teacher has a clear memory of the lesson to which the feedback applies. Second, a feedback checklist corresponding to the indicators and attributes does not inform the teacher of the specific behaviors that were observed and judged to be acceptable or not acceptable. Thus, if a teacher receives a negative rating for the student rapport attribute of the "Positive Learning Environment" indicator, she/he has no way of knowing which behaviors observed contributed to that rating.

Consideration also needs to be given to the use of mentor teachers in the feedback process. As mentioned earlier, teachers in Connecticut are encouraged to share their feedback results with a mentor teacher in order to get assistance in interpretation and guidance for improvement. During the pilot test in Connecticut, however, mentor teachers reported feelings of ambivalence or reluctance about participating in the evaluation of beginning teachers. Such feelings may have resulted because (1) the mentor teachers reported not feeling especially knowledgeable about the CCI and thus not well equipped to advise beginning teachers on either interpretation of the assessment results or ways to improve the areas found to be unacceptable, and/or (2) the mean number of times the mentor teachers reported observing beginning teachers is two, and thus they may not be familiar enough with the beginning teachers to recognize patterns of behavior that might illustrate adequate or inadequate performance. Should a classroom observation assessment be selected for use in California, a system of feedback should be developed that aids teachers in improving their performance. If mentor teachers are to be a part of this system, then they will need to receive training in the instrument and have time to observe the teacher to be of any real help.

Cost Analysis

We have used the experience and time associated with administering and scoring the current version of the CCI as a basis for providing some initial estimates of the costs of administering a California version of an observation system. We will outline the assumptions and basis for estimating the costs. It is important to view these as only general, incomplete estimates. To provide for more specific and complete estimates, it would be important to assess the feasibility of alternative methods for implementing the CCI. These could include varying the method used to administer and score the observation, and using alternative methods to allocate and absorb the costs of administration. For example, it might be possible to develop an observation system that reduces the scoring time from the four hours needed for the current version to one hour. Also, the CCI could be combined with other assessments such as interviews, assessment centers or written examinations in a manner that would affect the costs of administering each.

Assessor Time

Administering the current version of the CCI requires a trained observer or assessor to (a) prepare for and arrange for the assessment, (b) review the pre-assessment form, (c) conduct the pre-assessment meeting, (d) observe for 45 minutes to one hour, and (e) analyze the teacher's performance according to the ten indicators. The current analysis system requires as many as four hours to summarize and score an observation.

Allocating up to four hours for scoring and two hours for the other activities would imply that the assessment could be completed within six hours. Using an hourly rate of $20 per hour would cost $120 per observation for the assessor's time, which would account for the majority of costs. If the scoring time were reduced to one hour, the cost of assessor time would drop to $80 per observation.

Training Costs

The current version of the CCI requires a five-day training session and a two-day follow-up session. If we assume that each assessor could be trained and certified in this amount of time and that each would conduct 30 observations each year for five years, we could distribute the costs for training the person would be distributed over 150 observations. Reimbursing the assessors for the seven days of training at $20 per hour or $160 per day would add about $7 to the cost of each assessment.

Other Costs

Other costs would include those associated with telephone, duplication, postage, and travel where needed. Travel could be expensive in a state like California unless regional assessors were used. A regional system of assessors that involved little travel would minimize the cost. Placing an estimate on the costs of these activities or ingredients would depend in large part on the manner in which the system was ultimately designed and how costs were apportioned. Using a figure of $30 per assessment of these activities would assume only minimal travel costs, based on our experience from the pilot testing.

The above analysis results in the following cost estimate to administer a revised version of the CCI. The total cost could be as low as $117 per assessment if the observers could score the observation in one hour.

Assessor:	$120/assessment
Training:	$7/assessment
Other:	$30/assessment
Total:	$157/assessment.

The actual costs of implementing an assessment like the CCI and conducting multiple observations would depend on several factors as already mentioned. The costs for multiple observations would vary as a function of whether all candidates were observed the same number of times or whether only candidates who failed to demonstrate proficiency on early assessment(s) were subsequently observed. Additional costs would also include those for developing and managing the assessment system. But, the system design and the degree to which this type of assessment might be merged with other systems would affect the management and related costs. Estimates for these should be made after some of these alternatives are explored and specified.

Cost Summary

The experience from pilot testing a limited number of CCI assessments yields some early glimpses of the costs that might be associated with such an observation system. While the above analysis outlines costs for most of the ingredients that might go into a system, more refined estimates need to be made after the assessments that might be conducted are better defined and decisions on variations to the CCI (e.g. methods for scoring) are made.

Technical Quality

This section briefly discusses three technical issues related to the CCI -- development, reliability, and validity. No statistical data were available from either Connecticut or from the California pilot test.

Development

As previously discussed in the **Content** section of this chapter, the development of the CCI relied on a standard approach and included several major steps. The first was the identification of indicators of competent new teachers. Another was a review of the literature on effective teaching. Drafts of CCI materials, including indicators and attributes, were reviewed by national experts. During March 1988, Connecticut conducted a pilot test with 42 assessors and 36 new teachers in 27 school districts. Teachers, teacher trainers, administrators, and other experts have been involved in all phases of the development process.

Once the CCI was nearing the final draft stage, 1,582 Connecticut educators participated in a content validity study, reviewing the instruments in terms of relevance and importance. During 1988-89, Connecticut conducted a major field test involving 250 new teachers in 67 school districts. This included teachers in vocational-technical schools and addressed the generalizability issue across subject areas and grade levels. A bias review was also conducted and a formal standard-setting process was completed.

The development process used by Connecticut was sound. However, without additional information on the specifics of the steps used and individuals involved, little can be said about the quality of the effort. The available evidence suggests that the CCI was developed in a professional and technically acceptable manner.

Reliability

Several steps have been undertaken by Connecticut to help ensure a reliable assessment with the CCI. These include: (1) the training of assessors, (2) the selection of assessors who are experienced in teaching and in the subject area when feasible, and (3) the use of multiple observations of the same new teacher (six per teacher by different observers). In order to ensure consistent application and accuracy, Connecticut trainers review potential assessors on five areas -- completion of both sides of the "t-sheets," appropriateness of the data used for evidence of the defining attribute, the inclusion of comprehensive data for evidence, the writing of evidence in a way that specifically links data to the defining attribute, and the listing of specific examples of classroom behaviors, activities or circumstances. All of these procedures are designed to reduce the error in CCI results and thus promote its reliability.

However, no information was provided by Connecticut on other aspects of the CCI's reliability. Data that should be collected and reported include:

- o Inter-rater reliability -- two or more observers of the same lesson at the same time, including different types of observers (e.g., teachers vs. administrators);
- o Stability of ratings -- same teacher, same observer, different days;
- o Review by second observer -- of script, ratings and other documentation; and
- o Monitoring -- of rating patterns of each observer across several teachers to identify those observers whose ratings tend to be higher or lower on the average than other observers, *or* those who consistently rate certain indicators or attributes high or low compared to other observers, so that discrepant observers can be identified and retrained as needed.

In addition, the training system should include systematic reviews, even for observers who are not discrepant, to minimize the chances of their starting to drift and to maintain standards.

Validity

Several steps, as described under "Development," were undertaken during the development stage to ensure the validity of the CCI, including the identification of important indicators of new teacher competence, the literature review, the review of drafts, and the content validity study. Validity must be judged in terms of the use of the instrument. Validity is not inherent in the instrument itself. An instrument considered to be valid in Connecticut for teacher certification may or may not be valid for credentialing in California.

Little can be said about the validity of the CCI for California or its appropriateness for various subject areas, grade levels, students groups, and school/community settings based on the pilot test. If the CCI or any other high-inference observation instrument is field tested, California should conduct a validity study that considers the appropriateness of the instrument for these various settings, its relevance to a new teacher's job, the importance of each attribute and indicator in effective teaching *and* in protecting students from teachers who lack certain competencies, and the fairness of the CCI to new teachers in terms of their opportunity to acquire the skills being observed. New teachers, teacher trainers, mentor teachers, and teacher supervisors should be involved in a review of the validity of the instrument. They should also be asked about the clarity of the content and the process. Lack of clarity in either area will negatively affect both the reliability and validity of the assessment instrument.

Although reliability addresses the accuracy of the decisions made, based on the CCI ratings, a validity issue for California is the question of how much additional information is provided for use in making credentialing decisions. For example, are any decisions changed when the observation ratings are used in conjunction with other data already available (e.g., college grades, NTE scores, CBEST scores)? In those cases where different decisions are made once the ratings are considered, are the changes warranted? Will students be better protected from teachers who lack needed competencies to teach effectively if an observation instrument is used in conjunction with currently available information or other sources of information? Will teachers be more fairly assessed if additional information is available? Related to this issue is the question of how many observations are needed for each new teacher. Are six observations necessary or is there enough information after four observations to make decisions for the majority of cases? If the latter, two additional observations might be done only for the borderline cases. These questions require much more data and a much larger sample than was available in the CCI pilot testing, but must be addressed prior to adoption of this or any other assessment instrument for credentialing in California.

Conclusions and Recommendations

This section contains conclusions and recommendations regarding the CCI, organized into the areas of administration, content, format, and a brief summary. These conclusions and recommendations would likely apply to any high-inference observation instrument.

Administration of Assessment

The administration of the CCI assessment is very labor intensive, requiring nearly one professional person day per teacher. Seven observations in five days were deemed stressful by the assessors and independent observers; one per day is a more feasible workload, unless substantial travel time is involved. If a subject and grade-level match between the teacher and assessor is desired, the complexity of scheduling increases markedly, probably increasing the time required to administer observations because of greater assessor travel time.

The following factors seem to be key to smooth administration of the CCI in its present form:

- making and confirming arrangements with both principals and teachers regarding the time of the observation and the locations of the pre- and post-observation interviews;
- careful design of observation schedules for assessors, with no more than one observation scheduled per day;
- development of procedures for obtaining completed assessment materials from assessors in the field; and
- arrangements for storage of a large amount (at least 25 pages) of documentation per teacher.

Since the CCI assessment is administered and scored by the same person, the training of assessors is also a key factor to successful administration of the CCI. Through training, assessor candidates are taught the content of the assessment, as well as how to conduct and score the assessment. Current training consists of seven instructional days plus time to conduct practice observations. Training could be improved through the inclusion of more specific examples of written evidence and more time spent analyzing evidence from previously prepared scripts instead of scripting from videotapes. It is unlikely, however, that the training could be shortened considerably. Some modifications in the training would also be needed to accommodate the California context, reflecting the greater diversity of students, larger class size, and more frequent use of instructional aides. The training should conclude, as it does in Connecticut, with each assessor being required to exhibit a minimal level of proficiency in administering the assessment.

Assessment Content

Based on our observations and those of RMC staff, as well as information collected from assessors, teachers, and CCI rating sheets, we offer the following conclusions about the content of the CCI:

- Congruence of the CCI with the various California curriculum guides and frameworks is relatively weak. This is largely because (1) the CCI was developed in the context of the Connecticut curriculum; (2) it is a noncurriculum specific, high-inference observation system, and, (3) it is not designed to measure a teacher's knowledge of curriculum directly.

- Coverage by the CCI of the California Standards for Beginning Teachers varies. Coverage is particularly good for those standards which focus on student rapport, classroom environment, and student motivation, involvement, and conduct. Coverage is partial, however, for the majority of standards, and nonexistent for a few. Moreover, some standards partially covered, e.g., Curricular and Instructional Planning Skills, are difficult to measure using a classroom observation system.

- The job-relatedness of the CCI seems to be high because the assessment entails observing teachers actually teaching in their own classrooms.

- Overall, the content of the CCI does not seem too difficult for beginning teachers. Approximately 80% of the pilot test participants received passing scores (i.e., received an "Acceptable" rating on at least seven of the ten indicators).

- A variety of subjects, grade levels, community contexts, and instructional techniques were observed. The CCI appeared to focus on teaching abilities that are applicable in all K-12 instructional contexts. The appropriateness of the CCI for assessing teachers of classes with less emphasis on direct instruction and more emphasis on practice activity (e.g., physical education, band, vocational education) should be studied further.

- Analysis of the rating results by grade level (i.e., elementary, middle school, and high school) indicates that further pilot testing with a focus on grade-level matches between assessors and teachers may be useful and warranted.

- Subject-matter and grade-level matches between the assessor and the teacher observed might complicate administration considerably, but they would probably improve the instrument's ability to assess the appropriateness of content and lesson development.

- Although the creators of the CCI were sensitive to the issue of teaching diverse students and developed an assessment which focused on a teacher's interaction with all students, the CCI would need to be modi-

If the CCI is chosen for further development, a content validity study should be conducted in which California educators examine the instrument for the job relevance and relative importance of its indicators.

Assessment Format

The classroom observation format will be discussed at length in Chapter 6 and contrasted with the semi-structured interview and multiple-choice examination methods of teacher assessment. One strength of the CCI format is that its focus is not on a simulated performance, or on how a teacher says she/he would perform, or on a teacher's knowledge of how to perform, but rather on a teacher's actual performance in the classroom. In addition, because the teacher is observed in his/her own classroom, no special facilities are required for administration.

The format of the CCI goes beyond traditional observation systems in which an assessor checks off observed teaching behaviors. The CCI requires an observer to first script as much as possible the entire lesson observed, and then to document from this script the evidence supporting the existence or absence of the desired teaching behaviors. Such careful documentation greatly reduces the risk of an observer's subjectivity with regard to the teaching behaviors perceived and/or toward the teacher observed.

The CCI format also differs from other observation systems in that the actual observation is preceded and followed by interviews which are designed to (1) help the observer understand the instructional goals and classroom context which affect lesson design, and (2) give the teacher an opportunity to explain and justify changes in the original lesson design in response to unanticipated circumstances. The information provided in the two interviews and through the pre-assessment information form (which is completed by the teacher before the observation) allows the observer to conditionally evaluate teacher behaviors in light of differing instructional goals and classroom contexts. This observation instrument is superior to others used in teacher assessment because it focuses on the **meaning** rather than **frequency** of teacher behaviors.

Finally, the format of this assessment requires that the participating teachers receive complete information about the CCI, including descriptions of the indicators being rated, copies of all the interview protocols, and a sample completed copy of the pre-assessment information form. Based on the responses of the participating teachers and assessors to these materials and to the CCI format, we recommend the following:

- o In preparation for the CCI assessment, a teacher must be familiar with a large amount of material (i.e., the content of the CCI), prepare a lesson to meet CCI standards, and complete a pre-assessment information form. Therefore, we agree with the Connecticut assessors who believe that appropriate use of the CCI requires that teachers have access to help in preparing for the assessment.

- o If, as is done in Connecticut, mentor teachers are expected to help teachers prepare for the CCI, it is crucial that the mentor teachers (and/or others who give assistance) are well acquainted with the instrument, and free to observe the beginning teachers often enough to be acquainted with their usual teaching behaviors.

- Because both teachers and assessors expressed a desire that feedback be a part of the CCI process, the provision of feedback should be considered. Also, if the CCI is intended to serve as a guide for staff development as well as a requirement for credentialing, then the scope of the assistance needed by a beginning teacher to interpret the results (i.e., feedback) needs to be investigated.

Summary

If classroom observations are selected as a form of teacher assessment for credentialing purposes, the CCI could serve as a fully developed prototype. Reviews by California educators may suggest that alterations should be made in the indicators and standards, but the procedures for conducting the observation and methods of scoring appear to need no further development.

CHAPTER 4:

SEMI-STRUCTURED INTERVIEW: SECONDARY MATHEMATICS

CHAPTER 4:

SEMI-STRUCTURED INTERVIEW: SECONDARY MATHEMATICS

Developed by the State of Connecticut, the Semi-Structured Interview in Secondary Mathematics is a performance assessment designed to assess the competency of beginning secondary mathematics teachers. Through an interview format, the assessment targets a beginning teacher's knowledge in the subject area of mathematics, exploring a teacher's thought process as he or she makes instructional decisions for students.

Two versions of the Semi-Structured Interview in Secondary Mathematics have been developed by Connecticut. They are similar, but focus on two different topics: (1) linear equations, and (2) ratio, proportions, and percent. Each version, however, consists of the same five tasks:

(1) **Structuring a Unit:** A teacher arranges ten mathematical topics in a sequence that is appropriate for teaching the unit, explains reasons for the ordering based on training and experience, and discusses how the chosen sequence might affect student learning;

(2) **Structuring a Lesson:** A teacher explains how a lesson might be constructed on a topic represented by several pages of a textbook;

(3) **Alternative Mathematical Approaches:** A teacher is given alternative solution strategies for a problem, chooses the approach(es) to use to teach students, justifies the approach(es) selected, and discusses the relative advantages and disadvantages of each strategy;

(4) **Alternative Pedagogical Approaches:** A teacher is shown five alternative curriculum materials, selects the approach(es) to use to teach students, justifies the approach(es) selected, and discusses the relative advantages and disadvantages of each method; and

(5) **Evaluating Student Performance:** A teacher is shown samples of student work that contain an error in the solution, identifies the error made, and offers suggestions about remedial instruction for each kind of error.

Since the two versions differ only in the focal topic, they will be discussed together and will be treated as a single assessment format, referred to as the SSI-SM throughout the chapter.

The SSI-SM format combines two assessment strategies: the semi-structured interview and the assessment center. As an assessment strategy, semi-structured interviews provide opportunities for candidates to respond orally to a standardized series of questions about tasks that are presented verbally by an examiner who uses a script or interview schedule. This interview is semi-structured in that it allows the use of follow-up questions at the discretion of the assessor when a candidate's answer is judged to be unclear or incomplete. An assessment center strategy allows for simultaneous assessment of a number of candidates, all of whom participate in a series of exercises or tasks which might otherwise be administered to candidates individually. In the case of the SSI-SM, the assessment was organized so that a group of candidates rotated through the set of tasks, with each candidate completing a different subset of tasks in the same time period. The order in which candidates performed the tasks was purposely varied.

The SSI-SM was developed by the Connecticut State Department of Education (CSDE) for use in a three-tier assessment system that is designed to strengthen its teacher education program and improve the quality of its beginning teachers. As briefly described at the beginning of Chapter 4, this system includes the following assessments (each of which is administered at a different point in the beginning teacher's career): a minimal skills test in reading, writing, and mathematics; a multiple-choice examination measuring subject matter knowledge (secondary teachers are assessed in their area of specialty and elementary teachers are assessed with a custom-designed elementary education examination); and a classroom observation assessment that evaluates teachers on the essentials of effective teaching. Together, these instruments serve as important means for assessing a teacher's content knowledge and general pedagogical knowledge (or teaching skills), but they do not assess the intersection of subject matter knowledge and pedagogical skill -- or what Shulman describes as pedagogical content knowledge. Thus, in order to ensure that its teacher assessment program measures all essential components of teacher knowledge, the CSDE decided to develop and add to its system an instrument that would measure a beginning teacher's pedagogical content knowledge.

The resulting instrument was a semi-structured interview with a focus in mathematics. The subject area of mathematics was chosen for three reasons: (1) mathematics is generally acknowledged to have a more tightly defined knowledge base than such broader subject areas as social studies or language arts; (2) Connecticut State's math consultant was especially interested in developing an instrument of this kind; and (3) there is a relatively strong research base about teachers' cognitive processes in the area of mathematics.

Development of the SSI-SM proceeded as a collaboration between the CSDE and Gaea Leinhardt, an expert in the cognitive research on Mathematics teaching who has extensive experience working with teachers, as well as a background in testing and measurement. Throughout periodic stages of its development, the SSI-SM underwent review by educational researchers, curriculum experts, members of the State's Math Advisory Committee, and the interviewers who participated in the pilot study conducted in December 1986 and July 1987 involving 24 beginning and experienced teachers. Although the final instrument focused on math, the developers specifically designed it as a prototype to be generalizable across disciplines.

The administration, content, and format of the SSI-SM are discussed below. Following a discussion of the cost analysis and technical quality of the assessment, the chapter closes with a summary of conclusions reached, together with recommendations for further steps in exploring the feasibility and utility of the Semi-Structured Interview in Secondary Mathematics (or a similar instrument) in California teacher assessment.

Administration of Assessments

This section begins with an overview of the administration of the assessments, which is followed by a discussion of the logistics of administering the SSI-SM.

Overview

The SSI-SM was administered by six trained assessors from Connecticut, all of whom conducted the interviews during the week beginning May 10, 1989, at two different sites. The sample for this assessment was 20 secondary mathematics teachers. Table 4.1 contains information about the number of teachers assessed from each local pilot project, the assessment sites, and some of the characteristics (e.g., gender, grade level, ethnicity, and teaching experience) of the participating teachers.

Logistics

Logistical activities for this assessment included: (1) developing orientation materials for teachers and principals; (2) identifying teacher samples; (3) making travel arrangements for six trained assessors from Connecticut; (4) scheduling the test administration; (5) arranging facilities; (6) acquiring materials for the administration of each task; (7) arranging for the acquisition of evaluation feedback from teachers; and (8) arranging for district reimbursement for the cost of substitute teachers and payment to some teacher participants. Logistical arrangements are described in detail in the *Administration Report for Spring 1989.*

Teachers received a two-page description of the assessment by mail prior to the interview. Shortly before the interview began, they were given general information about the tasks and the purpose of the pilot test. Largely because this assessment had not yet been pilot tested and thus sample responses were not available, the information did not include a full range of descriptive sample material that is usually accessible to candidates. Providing a full range of descriptive sample material, however, is particularly important for this type of assessment, which departs dramatically in form from that of current California teacher assessments.

Concerns for due process and equal access have motivated most test publishers to provide candidates with orientation materials describing the examination's purpose, content, format, length and evaluation standards. Often sample assessment materials are made available. If this or a similar assessment were adopted, teachers would need timely delivery of materials with sufficient descriptive detail to allow for preparation and review prior to the assessment. Assessment orientation materials would need to be developed to provide this kind of information. Such materials might describe the purpose, format, and rationale for this new type of format, provide sample tasks and component questions, and discuss the type and range of potential topics and, most importantly, the criteria for evaluating candidates' responses. All current topics that might be assessed could be published and sent to all "registered" candidates a month before the assessment. Orientation materials for performance assessments such as essay

TABLE 4.1

SEMI-STRUCTURED INTERVIEW IN SECONDARY MATHEMATICS: PILOT TEST PARTICIPANTS

(Total Number of Teachers=20)

Descriptive Characteristics of Participants	Distributions of Participants
Teaching Experience	
First Year	13
Second Year	7
Teaching Level	
Middle Level	8
High School	12
Gender	
Male	11
Female	9
Ethnicity	
American Indian	1
Asian-American	1
Hispanic	2
White	16
Location of Teaching	
Fresno	12
New Haven	4
Oakland	4

examinations sometimes also recommend content and strategies that candidates might review in preparation for the assessment, and provide annotated examples of how responses are evaluated.

Based on the previous pilot experience, the tasks were refined and grouped into two sets so that each set of tasks took approximately equal amounts of time; the California experience was that the effort to balance the length of time of tasks was largely successful. Due to individual differences, there is almost always some variation in the length of time. Therefore, some arrangement must be made for smooth transitions between tasks. Procedures for handling especially verbose teachers who take longer periods of time need to be established. Teachers who do not communicate useful information in light of the scoring criteria can be prompted to finish; the more difficult decision is when to cut off the occasional teachers who take a long time to complete tasks because of superior breadth, depth, and detail.

Facilities required for this assessment included four interview rooms (one for each task) and one coordination room (for assembling before and between interviews) for every day of interviewing. Although FWL staff investigated a wide range of sites, we experienced severe difficulties in locating appropriate facilities with large numbers of small rooms. If the assessment were held on weekends or during the summer, vacant school or college classrooms could be utilized; for assessments held during the school week, similar problems in locating facilities can be anticipated.

The interviews were videotaped to provide a visual record for scoring the teachers' responses. (The scoring system was developed at a later date, precluding the option of scoring simultaneously with administration.) The use of videotaping equipment precipitated some disruptions and delays due to technical problems. Clearly, if videotaping continues as an assessment component, a technician needs to be close at hand. However, the necessity of videotaping rather than audiotaping should be further considered. To date, there seems to be little indication that any visual information is pertinent to evaluating responses or monitoring the assessors. Audiotaping is less technically demanding and more cost effective, in addition to preserving the anonymity of the candidates.

As with the other assessments, teachers were asked to complete a questionnaire asking their opinions of the SSI-SM in which they had just participated. Teachers were not asked to differentiate between the tasks for linear equations and those for ratios, proportions, and percents. Assessors also completed a feedback form on the final day of the assessment. They were asked to provide their perceptions of the adequacy of their training to administer the instrument, the logistical arrangements and facilities, the assessment format, the fairness of the instrument, and its appropriateness for assessing the teaching competence of new teachers.

Security

As with all assessments, the security of teacher evaluations is required. The extent of security necessary for the assessment materials is unclear. On the one hand, the answers to questions for each task are interrelated. One could not memorize isolated answers to questions, but would need to memorize an entire script. On the other hand, at one time, the state of Georgia included a semi-structured interview to gather information about a teacher's individual portfolio as part of its assessment instrument. The standardized questions allowed the development and teaching of standardized answers to beginning teachers which circumvented the effectiveness of the assessment, so the interview portion of the Georgia assessment was deleted. Before the adoption of this or any other semi-structured interview assessment, the robustness of semi-structured interviews with respect to development of standardized answers would need to be investigated.

Assessors and Their Training

Six trained assessors from Connecticut participated in the pilot testing. As was the case with the CCI pilot test (Chapter 3), the use of trained assessors from Connecticut rather than California assessors served to reduce the costs -- both in terms of time and money -- of the SSI-SM pilot test. Using trained assessors from Connecticut both reduced the time needed to coordinate the pilot test (e.g., no assessor recruitment or training necessary) and eliminated the costs associated with these two activities.

The six Connecticut assessors were all mathematics teachers with over five years of teaching experience. In Fall 1988, each participated in a one-day training session which included background information about the semi-structured interview, approximately two hours of lecture on methods of interviewing, and three hours of practice in administering interviews. Following the training, in November 1988, the assessors participated in a pilot study of the SSI-SM in Connecticut, administering one complete interview to 10 teachers. In preparation for the Spring 1989 pilot testing in California, a specialist in interpersonal communications gave the assessors refresher training in April 1989. The refresher training consisted of (1) roughly two hours of lecture on findings from the November pilot study; (2) a one-hour discussion of interviewing weaknesses identified through analysis of the November videotapes (e.g., probing tactics, establishing rapport with the candidate, and maintaining standardization across the interviewees); (3) a one-hour group discussion concerning changes made in the protocols; and (4) one-and-a-half hours of interviewing mathematics teacher education students using the new protocols. One important change that resulted from the November pilot was that each assessor was trained to administer a set of several tasks rather than a complete interview.

In this pilot test there were six assessors for four activity stations. Based on this experience, FWL has determined that four qualified assessors can adequately handle four stations. Additional assessors might be used for either training or coordination purposes, but are not needed for managing the assessment.

All of the assessors believed that their training had been adequate. Three assessors mentioned the importance of practice in administering the tasks, with one of them remarking that the training was inadequate without it. Two of these assessors also mentioned the work on probes as being useful. The only suggestion for improvement came from another assessor who would have liked more feedback on his performance.

Knowledge of the subject area is necessary to construct appropriate probes. Although our observations of the questioning indicated that the scripts prompted a fairly high degree of comparability, there was some unevenness in probing for additional detail or explanation. Analysis of pilot tapes suggests the desirability of such probes to better reflect the extent of the teacher's knowledge. However, the degree of probing can affect the rating of a teacher's response, with uneven probing affecting the fairness of the assessment across teachers. This is a dilemma which needs further exploration with particular attention to such issues as the variability of assessor probing, the conditions under which the variability occurs, and the implications of the analysis for selection and further training of assessors. Improved alignment between questions and scoring criteria would also reduce, but probably not eliminate, the necessity for probing. During meetings on the development of the scoring system, some interviewers suggested that explications of the scoring criteria should inform the development of guidelines for types and degrees of probing beyond the scripted questions.

Assessor training will be significantly altered if interviewers also serve as raters. Whether having the interviewer assume both roles will negatively affect the composure and/or performance of the candidates cannot be determined from this pilot test, but this issue will be a significant question for further pilot or field testing. The circumstances under which and the manner in which an assessor takes notes, probes, and reacts to candidates' responses will need careful attention.

Teacher and Assessor Impressions of Administration

Seventeen of the teachers felt that the arrangements were reasonable. Two specific suggestions for improvement were longer breaks and earlier notification.

Assessor comments focused on equipment and facilities. Several mentioned the importance of setting up the video equipment well in advance and instructing the assessors in its use. Two assessors commented that the use of hotel guest rooms did not contribute to a "professional" atmosphere. However, another assessor had the contrary impression, citing the desk-and-chair setting in a small hotel room as more professional than the use of a conference table in a mid-sized meeting room. One assessor was distracted by noise coming from the room next door.

Scoring

One purpose of the administration and especially the videotaping of the assessment was to allow the further development of the scoring system. (The videotaping allowed repeated viewing of an interview and the testing of different scoring methods.) The scoring approach has continued to evolve during the development of the semi-structured interviews. A team of consultants, Connecticut State Department of Education staff, and committees of Connecticut teachers have worked to augment the design of a scoring system that was pilot tested in Connecticut during the 1987-88 school year. The scoring system is in the final stages of development, but is not a completely developed prototype at this time.

One purpose of the continued development was to identify knowledge domains, key indicators, and quality criteria that might be applied across subject areas. The emerging scoring approach specifies three domains of expertise: Curricular/Content Pedagogy, and Knowledge of Students. There are currently two indicators or clusters of knowledge within each domain. To facilitate understanding of each indicator, it is augmented by elements, which are examples of specific knowledge, skills, or dispositions that comprise an indicator. The list of elements for each indicator is illustrative, and not definitive. The SSI-SM is scored at the indicator level.

Math Scoring Indicators and Indicator Elements

The specific indicators as of the fall of 1989 are as follows:

- CCI: Understands principles, skills, and concepts of the content area. This indicator is a test of content knowledge. Candidates are not expected to shine as outstanding mathematicians; however, inaccuracies in mathematics, gross or subtle errors in terminology, inappropriate representations of concepts should all point to weaknesses as a mathematician that will interfere with effective instruction. Elements include:
    - mathematics
    - mathematical terminology
- CC2: Understands mathematical interrelationships among topics and organizes content on the basis of the relationships. The mathematical concepts must be connected in a logical way based on interrelationships that create an appropriate curriculum for instruction. CCC2 addresses both the purpose and perspective of the content area. Elements include:
    - identifying prerequisite knowledge and skills
    - sequencing topics based on a mathematical perspective
    - grouping topics based on the mathematics addressed
    - identifying real world applications of topics
    - linking content to specialized skills (e.g., critical thinking)
    - linking content to a broader curriculum
    - analyzing texts, lesson materials, etc., as related to the broader curriculum
- CP3: Understands effective practices, successful approaches, and potential problems associated with mathematics instruction. This indicator is the core of the content-bound instructional knowledge. It asks whether or not the candidate is an effective teacher of mathematics capable of integrating his or her more general pedagogical skills and his or her knowledge of mathematics. This indicator should measure the common elements of "pure mathematics" and "pure pedagogy." CP3 should also address the richness of the teacher's instructional repertoire of content examples, analogies, materials, etc. Elements include:
    - examining alternative approaches to instruction on the basis of content
    - examining the relative importance of topics
    - examining the relative difficulty of concepts
    - anticipating problems all students will encounter

  - adjusting instruction based on mathematical context and practical considerations
  - identifying supplementary instructional materials
  - selecting instructional approaches that are appropriate to the instructional objectives
  - selecting instructional activities that are appropriate to the instructional objectives
  - demonstrating an instructional repertoire appropriate to the content area, including examples of concepts, effective analogies, multiple procedures for teaching concepts, representative analogies, and sound presentations

o CP4: Understands effective instructional practices that facilitate learning and are independent of the subject area. This indicator measures the candidates general pedagogical knowledge. Any evidence of sound instructional approaches that are independent of the content area should be credited under CP4 rather than CP3. Skills taught in a traditional teacher preparation program related to classroom management, lesson planning, lesson monitoring, routines and transitions, and general evaluation are represented in this indicator. Elements include:

  - structuring a lesson
  - providing clear opening and closing to a lesson
  - monitoring student understanding during direct instruction
  - monitoring time on task
  - maintaining routines to facilitate transitions from one activity to another
  - maintaining a sense of order in the classroom
  - encouraging student responsibility for their own learning
  - selecting appropriate grouping and other instructional strategies
  - fostering independence and interdependence of learners
  - evaluating student work (formal and informal)
  - providing feedback to students
  - evaluating instructional outcomes

o KS5: Justifies instructional practices and approaches on the basis of student background and interests. This indicator measures the extent to which the teacher considers the background, needs, and interests of his or her students, or students in general, in selecting instructional approaches that facilitate learning. One component of the indicator is the consideration of motivational strategies. Elements include.:

  - soliciting information about student background and interests
  - selecting lesson activities, presentations, and explanations that reflect student background and interests
  - designing instruction that considers the self-concept/self-esteem needs of students
  - connecting instruction to the real world experiences of students
  - building on the informal and intuitive knowledge of the students

- KS6: Justifies instructional practices and approaches on the basis of student abilities. Attention to student ability and the need to monitor and adjust instruction based on ability grouping, setting appropriate standards, etc. all contribute to this indicator. Elements include:
    - soliciting information about student abilities
    - selecting lesson activities, presentations, and explanations that reflect student abilities
    - modifying instruction to build on a student's existing mathematics knowledge base
    - developing alternate approaches to instruction for a given concept based on a range of individual student skills
    - identifying special approaches to instruction for highly capable/less capable students

Videotapes of teachers were viewed by the Connecticut development team and a number of Connecticut math teachers to identify instances of specific levels of performance. Once consensus was reached on these "marker tapes," the performances were compared to identify key distinctions between them. These distinctions were codified into descriptions of performances for each rating category. The marker tapes were then used to anchor the professional judgments of the scorers to the set of established standards.

Scoring Process

Although future plans call for the development of a scoring system which is implemented as the teacher is being interviewed, at present, all interviews are videotaped and then scored offsite. The current scoring system requires the scorer to view the videotape of a task and then record evidence in one of three columns representing the three scoring domains described above. Evidence consists of notations of appropriate or inappropriate statements about mathematical concepts, instructional techniques, or conceptions of students by the teacher. Once the viewing is complete, the scorer reviews the evidence recorded under each knowledge domain and codes the statements according to the indicators. At this point, the scorer may decide that the evidence has been misclassified as to domain, and reclassify it into a more appropriate domain. Proper classification is critical to reliable scoring. To assist the scorer in appropriate classification of evidence, elements of each indicator have been further delineated to guide the coding process.

For each of the indicators, the rater then uses three response characteristics to evaluate the quality of the candidate's explanations and justifications: the appropriateness of statements; the breadth of the repertoire; and the depth with which the candidate provides specific, reasoned examples. The rater weighs the importance of each of these criteria and evidence for them in deciding on a summary rating for each indicator. The summary ratings, in increasing order of proficiency, are: "insufficient," "marginal," "sufficient," and "proficient." The rater writes key evidence from the candidate's statements that support and explain the summary rating. These summaries of key evidence, in turn, can be used to defend the ratings and provide feedback to the candidate. To date, it has not been decided how ratings for indicators at the task level will be aggregated into summative judgments either at the indicator level across tasks or across the entire assessment for credentialing purposes.

There is general agreement among the scoring development team that the present configuration of interview tasks does not yield information sufficient to score all indicators. The tasks are to be analyzed to identify the indicators for which the questions elicit sufficient information to make reliable judgments. Further development work, by either adding questions to tasks or developing new tasks, is considered necessary before the assessment can be viewed as a completed prototype. Also, indicators addressing the ability to design summative evaluations of student learning and the capacity to reflectively evaluate one's own teaching are being considered as possible additions to the present list.

The struggle in developing this scoring system, or any scoring system for a performance assessment, has been to identify the types of responses that typify effective teaching and to specify criteria for evaluating the quality of the responses. Connecticut teacher committees were asked to consider the following questions in weighing alternative scoring approaches:

(1) Is the approach based on a theoretical rationale that explains how it characterizes effective teaching?

(2) Are the behaviors and criteria derived from empirical research?

(3) Are the behavioral and quality indicators descriptive and objective, not subjective?

(4) Is the language specific enough to be clear?

(5) Will behaviors and criteria generalize to other topics in the subject domain?

(6) Can ratings and supporting evidence provide constructive feedback?

Answers to these questions must precede decisions about the numerical range of the rating scale, the number of ratings, and the development of the training system. California may want to consider the Connecticut rating dimensions and criteria according to some of the questions above.

Discussion of Scoring System

Significant progress has been made on the development of a scoring system for semi-structured interviews. The Domain-Indicator-Element structure seems feasible not only for secondary math, but has high potential for serving as a prototype for scoring systems for interviews in other subjects. This system seems more useful than curriculum-specific, grade-specific or task-specific ones such as that of the SSI-EM (which will be discussed in the next chapter). However grade-specific, this favorable evaluation is based on professional judgments rather than any strong empirical evidence. Such evidence would consist of the development of parallel semi-structured interviews in other subjects.

The three domains seem to adequately cover the range of responses in a meaningful way. Whether the indicators are collectively adequate is *questionable*. It is not clear whether indicators are now comparable in breadth; it seems likely that more indicators in the Content Pedagogy scoring domain are needed. There is general agreement that at least two more indicators are needed to address both informal and formal evaluation and reflective learning from experience. However, the identification of indicators within the Content Pedagogy scoring domain is problematic. Currently two indicators addressing content pedagogy and "pure" pedagogy are used. Forming dichotomous categories covering these two areas is difficult in practice. An operational definition has been adopted to sort responses into separate categories. A response falls under "pure" pedagogy if a change in topic would not result in a different pedagogical decision; otherwise, it belongs under content pedagogy.

Discussions of potential revisions of tasks and questions suggest that the nature of the task affects the quality of evidence that can be generated for specific indicators. The problem might be resolved with the use of multiple matrix sampling, in which the tasks produce strong evidence individually, not only for selected indicators, but for all indicators when considered collectively.

The latest revision of the indicators was intended to make them more independent of each other, to facilitate classification of evidence. The degree of interrater reliability that is achieved (discussed in this chapter in the section on Technical Quality) will inform a decision as to whether future interviews will be scored by two or more raters. If interviews are not routinely scored by two raters, then it may be desirable to rescore the interviews of teachers who fail or who score near the passing standard.

The present scoring system poses a potential problem in that the summative scoring categories confound or combine scoring and passing standards in their use of "sufficient" and "insufficient" as rating categories. It would be preferable to have a rating system of clearly defined levels of performance that are defined independently of the passing standard. Decoupling standards from summative ratings would allow the state to raise the passing standard as teacher preparation improved in response to new credentialing requirements without completely redefining the rating system at close intervals.

Scorers and Their Training

The training system is also currently under development, using a holistic strategy akin to that of writing assessments. In writing assessments, the training consists of presenting numerous practice exercises on responses typical of each rating category. The length of the interview for each task makes this a more difficult problem than reading short essays. The identification and use of sample responses that merit a particular rating on a specific indicator are called "marker tapes." Marker tapes have made the scoring system easier to use and seem critical to reliable implementation. Future versions of the training system will probably need to continue the use of videotapes (or audiotapes) of candidates being interviewed about an entire task to illustrate rating decisions. Although the domains and indicators are identical for all tasks, scorers will need to be trained separately for each task with respect to definitions of the levels of performance corresponding to specific rating categories.

The training system should conclude with a qualifying set of exercises where trainees rate candidates' responses independently. If the trainee's ratings agree with ratings given by the expert validation committee, the rater qualifies to rate independently; if not, the rater must have more practice in applying the criteria. This step is seldom documented in performance assessments, but should be in order to establish the quality and credibility of the raters. Periodic checks or tests of scorers' agreement with prescored vignettes are also infrequent, but are recommended procedures for maintaining agreement levels and for preventing rater drift.

Assessment Content

The content of the assessment includes the knowledge and strategies to be measured and the types of tasks intended to elicit teachers' expertise. The tasks of the semi-structured interviews have been designed to represent significant, recurring activities that teachers engage in as they plan their instruction, present and adapt it, evaluate their students' progress, and reflect upon the effectiveness of their teaching. Ideally, the tasks should represent the range of topics that the candidate will be credentialed to teach, as well as a range of contexts and students. The tasks are intended to tap the beginning teacher's command of content, pedagogy, and knowledge of students. Task components tend to ask candidates to describe what they would do and to explain why they would do it.

This section discusses the following content-related aspects of the assessment:

o Congruence with California's curriculum framework and standards;

o Congruence with California's Beginning Teacher Standards;

o Job-relatedness of the content;

o Appropriateness for beginning teachers;

o Appropriateness as a method of assessment;

o Fairness across groups of teachers;

o Appropriateness across different teaching contexts

Congruence with California Curriculum Guides and Frameworks

The SSI-SM was developed by Connecticut for use with Connecticut teachers. FWL compared the tasks and the September 1989 version of the scoring system with the most recent California mathematics curriculum document, to see if they were congruent.

The topics chosen as the focus of the assessment are included in the strands, or groups of topics, in the secondary mathematics curriculum described in the 1985

Mathematics Framework for California Public Schools: Kindergarten through Grade Twelve (which will be referred to as the *Mathematics Framework*). Ratio and proportions are part of the number strand and are to be taught by the end of the eighth grade. Linear equations are part of the algebra strand and may be introduced in one of two ways. The first possibility is in a ninth-grade algebra class as part of the preparation for higher level mathematics. The second possibility, for students who do not intend to pursue advanced mathematics, is in the first year of a two-year sequence that equips students with the basic mathematical knowledge needed in a technological society.

The *Mathematics Framework* lists five major areas of emphasis: (1) problem solving (by which is meant the ability to solve applied problems, and not the routine application of algorithms to textbook problems); (2) calculator technology; (3) computational skills; (4) estimation and mental arithmetic; and (5) computers in mathematics education. The SSI-SM addresses computational skills and computers in mathematics education. Problem solving was the focus of one element of a scoring criterion, though no questions specifically asked about strategies to teach problem-solving; calculator technology, estimation and mental arithmetic were not addressed at all. The specific ways in which the emphases were included in the assessment were:

- o Although the scoring system is still under development, practice scoring sessions observed by FWL staff included discussions of scoring the domain of Content Pedagogy which clearly indicated that the emphasis in scoring is on teaching students mathematical concepts and reasoning instead of memorization of mathematical algorithms. This is consistent with the discussion of computational skills in the *Mathematics Framework.*

- o One of the teaching techniques in **Alternative Pedagogical Approaches** was the use of computer software. Teachers were asked to relate the advantages and disadvantages of using the particular software compared to alternative strategies that might be used to teach the same concept. So while a software package was included, teachers were not asked to discuss the broader range of potential uses of computers in the classroom.

- o One of the response criteria (i.e., appropriateness) for the scoring domain of Content Pedagogy includes evaluation of the teacher's modeling of problem-solving processes and operations, but there are no specific tasks or questions which would direct a teacher to explain how she/he would do so. To fully address this *Mathematics Framework* emphasis, tasks and component questions might explore the teachers' range of techniques for promoting student's problem-solving strategies.

In addition to the five major areas of emphasis, the *Mathematics Framework* emphasized the following characteristics in terms of the delivery of instruction in mathematics: teaching for understanding, reinforcement of concepts and skills, problem solving, situational lessons, use of concrete materials, flexibility of instruction, corrective instruction/remediation, cooperative learning groups, mathematical language, and questioning and responding. About half of these areas were elements in the SSI-SM scoring indicators. The tasks and component questions directly addressed teaching for understanding, flexibility of instruction, corrective instruction/remediation, and mathematical language. Teaching for understanding, as discussed previously, was at the heart of the Content Pedagogy domain in the scoring system. One question in **Constructing a Lesson** specifically asked how and why a teacher might foster problem solving or critical thinking during the lesson. Flexibility of instruction was addressed by questions in **Constructing a Lesson** and **Alternative Pedagogical Approaches** that asked about "highly" and "less" capable students. Corrective instruction/remediation was the focus of **Evaluating Student Performance.** The teacher's correct use of mathematical language and concepts was one of the indicators of the Curriculum Content domain of the scoring system.

While identifying real world applications, an aspect of situational lessons, is one element of one of the Curriculum Content indicators, there are no questions or activities which would directly cue this type of response. The same was true for fostering interdependence of learners, an element of the Content Pedagogy scoring domain.

Table 4.2 summarizes the extent of coverage of the *Mathematics Framework*. The tasks as they are presently constituted address in depth computational skills, teaching for understanding, corrective instruction/remediation, and mathematical language. Although the scoring system as it presently stands has substantial coverage of the content and instructional emphases of the *Mathematics Framework*, the tasks themselves must be modified or, in some cases, redesigned to more directly address the remaining emphases to collect enough information to enable the scorers to make judgments for all emphases of the *Mathematics Framework*.

Extent of Coverage of California Standards for Beginning Teachers

FWL compared the SSI-SM tasks and September 1989 version of the scoring criteria with the 11 California Beginning Teacher Standards that student teachers are expected to attain when they complete California teacher preparation programs. The standards are composed of a general statement describing the competency together with factors which illustrate the subcomponents of the competency. Each standard is discussed separately. (Listed below are brief descriptions of Standards 22 through 32 with each standard defined in italics, along with descriptions of how the CCI indicators correspond to the standards), along with a discussion of how the tasks in the SSI-SM address each standard.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* None of the tasks in the SSI-SM address this standard. Indeed, except for the "clearly stated expectations regarding student conduct" specified by the standard, the other factors in the standard address teacher behavior when interacting with students. This might be difficult to simulate in an interview situation, compared to observing actual teaching. Questions could be developed which take one of two approaches: (1) ask a teacher to explain how they judge and evaluate their classroom environment and rapport with their students; or (2) require a teacher to evaluate and offer suggestions for a hypothetical class. However, the relationship between teacher responses to these tasks and their observed ability to establish rapport is likely to be slight.

TABLE 4.2

COVERAGE OF THE CALIFORNIA MATHEMATICS FRAMEWORK BY SSI-SM

Content	Method of Coverage	Extent of Coverage
Areas of Emphasis:		
Problem Solving	-An element of an indicator; requires development of tasks or questions to assess fully.	Partial
Calculator Technology	-Would require development of new tasks or questions.	None
Computational Skills	-Major focus of tasks and questions; implicit in scoring criteria and could be strengthened.	Partial
Estimation and Mental Arithmetic	-Would require development of new tasks or questions.	None
Computers in Mathematics Education	-A software program of a series of pedagogical approaches compared and constructed.	Partial
Delivery of Instruction:		
Teaching for Understanding	-Implicit in tasks; major theme of indicator.	Full
Reinforcement of Concepts and Skills	-Not directly addressed by tasks or questions; could be scored under an indicator.	None
Problem Solving	-Not directly addressed by tasks or questions; an element of indicator.	None
Situational Lessons	-Not directly addressed by tasks, questions, or indicators.	None
Use of Concrete Materials	-Not directly addressed by tasks, questions, or indicators.	None
Flexibility of Instruction	-Focus of questions in two tasks; breadth contributes to rating for all indicators; major theme of indicator.	Full
Corrective Instructions Remediation	-Focus of one task; an element of indicator.	Partial
Cooperative Learning Groups	-Not addressed by tasks, questions, or indicators.	None
Mathematical Language	-Implicit in tasks; major theme of indicator.	Full
Questioning and Responding	-Not addressed by tasks, questions, or indicators.	None

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies, activities, materials and assessment plans that are well defined and coordinated with each other.* This is the focus of the task, **Constructing a Unit**, and is partially addressed by the task of **Constructing a Lesson**. **Constructing a Unit** requires each teacher to order mathematical topics in a unit according to the best way to teach them. The appropriateness of the ordering and the teacher's explanation of the reasons underlying the ordering serve as evidence for two scoring domains: Curriculum Content and Content Pedagogy. (The measurement of the teacher's understanding of the selected mathematical topics and their interrelationship is beyond the scope of this standard, but such content knowledge is necessary to plan effective instruction.) **Constructing a Lesson** asks a teacher to plan a lesson on a given topic; however, only the single lesson and not the preceding or following lessons are described, so there is no opportunity to judge the coordination or development of a series of lessons, so the SSI-SM is only partially congruent with this standard.

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities and materials that are appropriate for students with diverse needs, interests and learning styles.* This standard is addressed to some extent by the tasks, **Constructing a Lesson** and **Alternative Pedagogical Approaches**, both of which ask questions about altering choices for "highly" and "less" capable students. The scoring domain, Knowledge of Students, has one indicator which specifically addresses adjusting instruction for students of different abilities and one which includes designing instruction to reflect student background and interests. There are, however, no questions which directly address the latter. Similarly, building on prior student learning is an element of the scoring domain of Content Pedagogy, but there are no questions which elicit information about how teachers plan to do this. Diversity of interests beyond academic interests and the use of a variety of approaches and materials that are free from bias are addressed neither by the SSI-SM tasks nor by the scoring system. The addition of questions or vignettes or the development of a new task would be necessary to fully address this standard.

Standard 25: Student Motivation, Involvement and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* One question in the task of **Constructing a Lesson** asked how students would be actively involved during the lesson. The response to this question would most likely yield information that could serve as evidence for the indicator addressing motivation of students in the scoring domain, Knowledge of Students. The task of **Alternative Pedagogical Strategies** also asked the teachers to take into consideration the students' needs and interests. While "monitoring time on task" and "maintaining a sense of order in the classroom" are elements of the scoring domain of Content Pedagogy which address this standard, no questions or tasks directly elicit information to assist a scorer in making judgments about the teacher's competency for these elements. Equitable treatment of students is not addressed by the SSI-SM. To fully address this standard, a new task would need to be developed, e.g., either vignettes of student misconduct or questions eliciting a description of a teacher's student behavior management system.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* The discussion of this standard addresses the linguistic complexity and nonverbal aspects of a teacher's communications with students. Two of the SSI-SM tasks, **Evaluating Student Work** and **Constructing a Unit**, might yield information on teacher explanations of concepts which would be scored under the scoring domain of Content Pedagogy. Some aspects of a teacher's presentation during the interview, e.g., clarity of explanations, the spontaneity and organization of his or her responses, and degree of enthusiasm, might serve as a crude proxy for his or her presentation skills in the classroom. However, the degree of relationship between a teacher's behavior during the interview and behavior in the classroom interacting with students, especially at the elementary level, would need to be investigated before using interview behavior as a proxy with any confidence.

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* Although **Evaluating Student Performance** focuses on the remediation of student errors, SSI-SM questions do not address the setting of high standards for achievement, ascertaining prior attainments, and designing and interpreting both formal and informal means of evaluation. A teacher might volunteer information addressing these factors in this standard; any such information would probably be scored under one of two scoring domains, Content Pedagogy or Knowledge of Students, with the exact scoring depending on the nature of the information. Another task would need to be developed to fully address this standard.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* Student outcomes are not directly addressed by the SSI-SM, nor could they be addressed directly in an interview format. However, the thrust of the scoring domain of Content Pedagogy is whether the teacher is laying a cognitive foundation that enables the student to achieve understanding of mathematical concepts and their interrelation. Content pedagogy is both one of the three scoring domains, and an indicator within the domain. The ability to use subject-specific content pedagogy is strongly assessed, but the degree to which one believes that cognitive outcomes of teaching are assessed depends on the confidence that one has in the links between content pedagogy and student outcomes. Lay audiences, including legislators, may desire more direct evidence of cognitive outcomes of teaching than are possible in a semi-structured interview format.

Standard 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward the subjects learned, the students themselves, and their capacity to become independent learners.* The encouragement of positive interaction among students and the provision for independent learning experiences is not addressed by any of the tasks in the SSI-SM. Student motivation was discussed under Standard 25.

Standard 30: Capacity to Teach Cross-culturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural, gender, linguistic and socioeconomic differences.* This is not addressed by the SSI-SM. Adaptation of the tasks or development of new tasks would be required to address this standard. One possibility is to provide more specific information about the classroom contexts for which the tasks are to be performed, and add questions asking how the context influenced the candidate's decisions.

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* Although this standard addresses student teaching experience, it can be construed to mean that teachers should be prepared to teach courses spanning the curriculum covered by the teaching credential. The SSI-SM does not do this; one possible revision would be to revise the tasks to address differing topics which occur at various points in the curriculum.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and develops professionally through self-assessment and collegial interactions with other members of the profession.* Neither respect for students and their ideas nor relationships with other teachers are addressed by the SSI-SM. This would require development of an additional task.

The extent to which the SSI-SM covers the California Standards for Beginning Teachers is summarized in Table 4.3.

Job-relatedness

Teachers were asked their opinion of the assessment's job-relatedness. Fourteen of the 18 teachers who completed the evaluation feedback form felt that all of the major tasks were relevant; three teachers did not, and one did not respond. One teacher with a positive response stated:

> *Yes, everything [in the SSI-SM] determines how successful my teaching is.*

A few teachers responded positively, but qualified their answers. One teacher, for example, noted:

> *Relevant yes, but [the SSI-SM tasks] miss the critically important areas of classroom management and control.*

Of the three teachers who did not feel all the major tasks were relevant, one teacher specifically criticized the emphasis on remediation of an individual student error on a single problem in **Evaluating Student Performance.** The teacher indicated this task is not realistic, given the large class size in California. Another teacher remarked, "Teaching is also a function of the students in your class." FWL staff interpret this teacher's comment as pointing out that the assessment does not capture a teacher's ability to tailor a lesson to a particular group of students with which the teacher becomes increasingly familiar over the school year.

All the CT assessors felt strongly that new teachers need the skills and knowledge that are reflected in the assessment to perform competently as a new teacher.

TABLE 4.3

EXTENT OF COVERAGE BY THE SSI–SM OF CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Standard	Method of Coverage	Extent of Coverage
22: Student Rapport and Classroom Environment	–Not covered.	None
23: Curricular and Instructional Planning Skills	–Focus of two tasks.	Partial
24: Diverse and Appropriate Teaching	–Covered by questions in two tasks. Breadth of content pedagogy and ability to design instruction taking students' ability and interests into account are major scoring components, though more questions should be added to fully assess abilities in this area.	Partial
25: Student Motivation, Involvement, and Conduct	–Covered by questions in two tasks and two elements of scoring indicators.	Partial
26: Presentation Skills	–Not directly covered by tasks, questions, or indicators.	Partial
27: Student Diagnosis, Achievement, and Evaluation	–Partial focus of one task.	Partial
28: Cognitive Outcomes of Teaching	–Not directly covered.	None
29: Affective Outcomes of Teaching	–Not covered.	None
30: Capacity to Teach Crossculturally	–Not covered.	None
31: Readiness for Diverse Responsibilities	–Not covered.	None
32: Professional Obligations	–Not covered.	None

Appropriateness for Beginning Teachers

The appropriateness of the SSI-SM content for beginning teachers was to be evaluated in two ways: (1) the perceptions of SSI-SM teachers, assessors, and other observers and (2) the performance of teachers on the assessments.

Perceptions. When asked whether the mathematical topics and concepts chosen for the assessment were appropriate for demonstrating their teaching skills, 16 of the teachers replied affirmatively. One teacher commented:

> *Yes, the topics were basic enough so that even if you haven't taught the lesson, the topics were appropriate.*

Another teacher concurred:

> *I haven't had to teach ratios/proportions, but probably will some day.*

One teacher, however, had exactly the opposite opinion. He found the assessment to be inappropriate because he had not taught linear equations to his seventh graders in the depth that he felt was required to respond to the questions.

Another teacher stated that the topics and concepts chosen for the assessment were fair, but that the assessment "failed to really challenge."

The CT assessors also believed that the subject matter content and tasks were appropriate means of assessing new teachers.

Eleven of the teachers believed that they had sufficient opportunities to acquire the skills needed to respond to the tasks, six did not, and one was not sure. Of the 10 teachers identified as being in their first year, six believed the tasks were appropriate for beginning teachers, two did not, and two gave qualified answers.

One teacher acknowledged the relevance of the tasks, but did not feel that his education had prepared him to perform the tasks competently. Another teacher implied that knowledge of how to adjust a lesson for a gifted or slow class depended upon experience teaching those types of students.

Two second-year teachers indicated that they needed the second year of experience to respond well to the questions. One teacher explained that if he had been asked the same questions when he was a brand new teacher, his answers would probably have been more idealistic and less likely to reflect "the reality of the school."

All but one of the CT assessors felt that new teachers would have had the opportunity to acquire the skills and knowledge needed to respond to the assessment tasks. Two suggested that if many teachers were having a problem, then that would reflect inadequate preparation programs for mathematics teachers at the secondary level.

The dissenting assessor thought that the tasks were appropriate, but that some new teachers might not have had first-hand experience teaching the particular topics that were the focus of the assessment:

> *Some teachers could (prior to the interview) not be at this point in their curriculum. These topics are usually taught in two different semesters at the middle-school level. Ratio and proportions might not be taught in the high school position of a first-year teacher. However, I think that both stimuli are appropriate. Student-teaching experience might provide background on these topics.*

Performance on assessment. The scoring of the tapes occurred in the later stages of writing this report, so no data on the performance of the teachers can be used to judge the appropriateness of the assessment for beginning teachers. The results and analyses of the scored interviews will be reported separately.

Because the teachers had little or no information about the tasks and scoring criteria, however, the results of the pilot testing will be incomplete at best. We will not be able to estimate what teachers might have done or can do when they have adequate information, well in advance of the administration, about what is expected and how it will be judged. Decisions about the generic features of the tasks, materials, questions, probes and scoring criteria should be made prior to future field tests.

Appropriateness across Contexts

Teachers were also asked whether the assessment is fair to teachers working in differing teaching contexts. Fifteen of the 18 teachers felt that the semi-structured interview approach was appropriate for teachers in varying contexts (e.g., across grade levels and subject areas, across various student groups, and in different school/community settings), two did not, and one did not respond. Appropriateness across grade levels, subject areas and diverse student groups are discussed below.

Grade level and subject area. This assessment was specifically tailored for secondary teachers of mathematics, with a particular focus on the topics of linear equations and ratio, proportions, and percent. Although the majority of teachers felt the assessment was appropriate across grade levels and subject areas, two teachers had different perspectives. One teacher stated he did not feel he had enough experience teaching linear equations to his seventh-grade class to fully respond to the questions on this topic. The other teacher with a viewpoint different from the majority felt that this type of assessment did not seem appropriate for all subject areas, commenting, "The assessment was more appropriate for objective subject-matter courses like math and science than for social studies classes."

Diverse students. Because of various constraints on the selection of the sample for this pilot test (e.g., availability of secondary math teachers in the same geographical proximity), not all teaching contexts were represented. Urban areas were overrepresented, and small cities and rural towns were not represented at all. Moreover, at least two of the teachers taught inner-city students in a context where the high-ability students were those who scored at the sixtieth percentile on achievement tests. None of the participating teachers, however, commented that the assessment was inappropriate for teachers in urban settings, of students of different ability levels, or of any other diverse student groups.

Two of the six CT assessors also believed that the assessment was fair to teachers of diverse student groups, with one citing candidates who mentioned "their personal experiences with ESL children and alternate school settings." The other four assessors, however, did not believe that the assessment yielded any information about the ability of a new teacher to work with diverse student groups.

Fairness across Groups of Teachers

All of the responding teachers felt that the assessment is fair among teachers of different gender and ethnic groups. It should be noted, however, that only three minority teachers were assessed. The CT assessors also believed the assessment to be fair across teacher groups.

Appropriateness as a Method of Assessment

As an assessment method, the strength of the semi-structured interview is that it can measure teachers' awareness of and reasoning about their cognitive strategies. Teachers can describe what they know and explain how and why they would apply their knowledge in a variety of situations. Unlike selected response formats, ranges of responses and interpretations are possible and acceptable. However, semi-structured interviews share the challenges of all performance assessments -- the standardization of tasks and questions, documentation of candidates' responses, and the application of explicit, uniform evaluation criteria for assessing performance.

Assessment Format

This section discusses the clarity, timing, and tasks of the SSI-SM, and summarizes the comments made by teachers concerning feedback on their performance. Eighteen of the 20 teachers who participated in the pilot test provided written input on the SSI-SM. This section is based on their comments as well as those received from the assessors and the observers from FWL and RMC.

Clarity of Assessment

Prior to the pilot test, teachers received a description of the assessment and the five tasks to be performed. Sixteen teachers reported that the written materials mailed to them were helpful; one did not find them helpful; and one qualified a positive response. Twelve teachers found the oral overview before the assessment to be helpful, while two teachers responded explicitly that it was not. The remaining four teachers gave varying responses, ranging from "N/A" to "A little" to a comment that the overview was repetitive. Suggestions for improving the orientation materials were given in the section on "Logistics."

All but one of the 18 teachers believed that the directions were clear; the other teacher commented that the second interview was easier because "you know what to expect." Fifteen teachers found the questions and tasks understandable; three did not. One teacher found the first task (evidently, "Structuring a Unit") confusing because it was "difficult to view each topic as being on the same level."

None of the CT assessors reported the need to make changes in the procedures or content of the tasks to accommodate teachers. The use of probes according to training guidelines seemed to enable the assessors to adequately adapt the tasks to different teachers.

As discussed earlier, the questions in the tasks and the scoring criteria did not always align well, due to the development of the scoring system after the administration of the assessment. The questions and tasks need to be redesigned to more fully elicit evidence to be scored for specific indicators.

Timing

The schedule for the SSI-SM allotted approximately the same length of time for each task, and specified the time the teachers were to move from one task to another. Assessors were asked to limit the time a teacher took to prepare for a task, although the time allowed was considered to be ample. There were individual differences in the length of time taken to complete a task, depending on the average length of explanations a teacher tended to give and the amount of probing needed. If teachers did not complete a task in the allotted time, often they were allowed a little additional time, although the assessor was instructed to try to move the teacher along. If teachers finished a task early, the coordination room was available to gather and take a break.

Fifteen teachers found the timing satisfactory; two did not; and one teacher did not respond to the question. One teacher mentioned specifically the flexibility in timing as an asset. This same flexibility in timing was considered a liability, however, by an assessor who mentioned the difficulty of taking a break or having a relaxed lunch when the interviews did not begin and end on time.

The sequencing of the tasks also needs further investigation. **Structuring a Unit** and **Structuring a Lesson** are more complex and time consuming than other tasks. Issues of fatigue and conceptual continuity should be considered to see if alternative task orders affect performance.

Nature of the Tasks

In the SSI-SM, candidates respond to stimulus materials that are intended to simulate the materials that teachers actually encounter. Yet some of the tasks include atypical materials. For example, **Structuring a Unit** presents concepts and topics on separate task cards; **Structuring a Lesson** presents pages from a textbook, but no array of supplementary materials, such as the teacher's guide; and in **Alternative Pedagogical Approaches**, the approaches are only sketchily described. California educators may wish to consider whether the type and range of resources in the SSI-SM represent the materials that California teachers are encouraged to use in organizing and presenting instruction in mathematics.

Teacher Preferences about Feedback

When asked what type of feedback they desired, six of the 18 teachers specifically mentioned information on their own strengths and weaknesses as reflected by their responses to the SSI-SM. Three wanted information either about scoring criteria or what the assessment was looking for in specific questions. Two wanted information so they could compare their own responses with those of others. Three teachers specified written feedback, two written or oral, and one wanted to watch the videotapes and then discuss his/her performance.

As to who should provide the feedback, two teachers suggested the interviewers, two others desired feedback from a committee, one teacher mentioned a master teacher, and another teacher recommended someone other than the teacher's supervisor.

Cost Analysis

SSI Cost Estimates

We can use the experience of administering the SSI-SM pilot tests as a basis to estimate the costs of implementing a semi-structured interview as a credential requirement in California. To review, the SSI-SM was administered in a setting in which four teachers rotated among four assessors in a half-day interview. Thus, four assessors could administer eight half-day interviews in one day. Assessors also would need some time to prepare for the interview and to summarize their notes and evaluations. The scoring system for the SSI-SM was not developed such that the interviewers could score and evaluate the interviews in the pilot test. To minimize costs, it would be helpful to have the interviewers evaluate and score the teacher interviews. Assuming that the interviewers could conduct the interviews and score them by allowing an additional two hours for preparation and evaluation, it would take approximately five hours of assessor time to complete the interview and score a half-day interview. Using the rate of $20/hour from our pilot testing yields the cost of $100/half day interview for the assessor time.

If we use the same training cost assumptions of $7/assessment that were used for the CCI, and the other costs for phone, postage, etc. of $30/assessment, the interviews would require approximately $137/half-day interview for each teacher.

Again, caution should be used in interpreting these figures since final costs will depend upon the actual requirements of the assessment, and of the system within which the interviews are placed. Furthermore, this analysis makes no assumptions about the manner in which the costs would be supported, e.g. charged to teacher candidates, supported by district or other staff, or supported by state agencies.

Technical Quality

This section discusses three aspects of the technical quality of the SSI-SM: development, reliability, and validity.

Development

As described earlier, the development of the SSI-SM has been based on both theory and research. Early versions of the interview protocols were reviewed by groups of researchers and practitioners. The scoring system is in the final stages of development; once it is completed, the completed assessment package will be reviewed by experts in the fields of mathematics, teaching and measurement to evaluate the assessment as a whole and to consider its strengths and weaknesses in relation to alternative approaches. While the SSI-SM is nearing the final stages of development, revision of the tasks to align them more closely with the scoring criteria will necessitate at least one further round of them pilot testing before the assessment can be considered ready for field testing.

Reliability

The consistency of the SSI-SM needs to be examined in several respects: (1) consistency of a teacher's performance across tasks; (2) consistency of a teacher's performance across topics; (3) consistency of a scorer across tasks; and (4) the internal consistency within tasks and topics. Pilot test data for the SSI-SM provide some initial information on the consistency of the SSI-SM and its scoring system. Appendix A contains a brief summary and display of data on the internal consistency and interrater reliability of the SSI-SM pilot test data. Basically, the information on interrater reliability suggests moderate agreement among the raters on the initial independent ratings and consensus ratings after raters discussed and revised their ratings. The percent agreements for the ratings are given in Table 4.4 on the following page. Nearly all ratings were within one point. Over half the ratings were exactly the same. Improvements in these agreements can undoubtedly be achieved with further development in scoring and training of raters. But, they do support that it is possible to achieve agreement among raters in an interview assessment such as the SSI-SM.

The pilot test information is not sufficient to judge the degree to which reliable formative feedback might be given within indicators, tasks or topics. However, the internal consistency on the Total Score provides some evidence that assessments like this can produce reliable decisions about individual candidates.

In summary, the developmental nature of the SSI-SM and small numbers of cases on which scores are available limit making any specific conclusions about its psychometric qualities. However, the agreements achieved among raters and the internal consistency that was exhibited suggest that assessments using this approach have the potential to achieve the consistency needed to provide reliable information on individual candidates. Realization of this potential would depend on further developments and refinements to the interview and scorer system and training.

TABLE 4.4

*PERCENTAGE OF RATER EXACT AND ADJACENT AGREEMENT BY TASK AND TOPIC PAIRS

Pairs	Percent Exact	Percent Adjacent
Pair 1	54	90
Pair 2	52	90
Pair 3	58	96
Pair 4	60	98

Pair 1 – Task 1, Linear Equations
Pair 2 – Task 3, Linear Equations
Pair 3 – Task 2, Ratio & Proportions
Pair 4 – Task 3, Ratio & Proportions

*This table represents mean independent rater agreements across the five indicators, across the ten examinees. Exact agreement means that each rater assigned the same rating as their rating pair. Adjacent agreement means that each rater was within one point of their rating pair on the assigned rating. Please note that the following analysis is tentative and has not as yet been verified by the SAS analysis.

Validity

Earlier versions of the SSI-SM have been subjected to some forms of judgmental validity through reviews by Connecticut mathematics educators. More content validation, including comparison of the same teachers using differing assessment formats, is indicated. These investigations should also include various forms of empirical validity, such as whether the assessment discriminates between beginning and experienced teachers, and between beginning teachers who are identified as more or less effective by other means.

Conclusions and Recommendations

This section contains conclusions and recommendations regarding the SSI-SM, organized into the areas of administration, scoring, content, format, and a brief summary.

Administration of Assessment

The SSI-SM is very labor intensive; at the present time, administration and scoring require one day per teacher. If on-line scoring is developed and found to be feasible, the overall time would be reduced slightly, but our experience with the CCI suggests that forming and documenting judgments takes a considerable amount of time.

Factors which are key to smooth implementation of the SSI-SM include:

- availability of appropriate facilities (which are often difficult to locate);
- development of clear orientation materials for teachers, including descriptions of the tasks and scoring criteria;
- organization of the assessment so all tasks take approximately equal amounts of time and only a minimal number of transitions between tasks are needed;
- if assessments are videotaped, arrangements for a technical consultant and familiarization of assessors with the equipment; and
- recruitment of assessors and scorers who are knowledgeable about mathematics, mathematics pedagogy, and student characteristics.

The cost of administering this assessment could be reduced by substituting audiotaping for videotaping as the form of documentation.

Scoring

The scoring system of the SSI-SM holds great potential for a multidimensional assessment of teaching competency. Its strengths include broad applicability across tasks, topics, and teaching styles and philosophies. Furthermore, the system's focus on three broad domains of teaching (Curriculum Content, Content Pedagogy, and Knowledge of Students) makes it likely that it will be suitable for semi-structured interviews in other subject areas. The latest pilot test of the scoring system suggests that despite reliance on professional judgment rather than checklists of observable behaviors, a high degree of interrater reliability can be obtained.

Before the SSI-SM scoring system is adopted, however, the following aspects need improvement:

- greater alignment between questions and indicators; and,
- redevelopment of indicators within the Content Pedagogy domain so that indicators are both comparable in scope and representative of all significant competencies falling within that domain.

The feasibility of simultaneously scoring and administering the assessment should be explored, including the identification of problems in combining the two roles and possible effects on interaction between the interviewer and the candidate.

Assessment Content

Our observations and information collected from assessors, scorers, and teachers participating in the pilot test suggest the following conclusions about content:

- The assessment content is in line with the philosophy of the Mathematics Framework. Congruence of the tasks is good, though not complete, with respect to the areas of emphasis and characteristics of delivery of instruction.
- The two topics constituting the content of the SSI-SM did not reflect the diversity of topics in the secondary mathematics curriculum.
- Coverage of the California Standards for Beginning Teachers is poor, but could be improved by refining current tasks and developing several new ones. Some standards that address teacher-student interaction or student outcomes could only be indirectly assessed.
- Questions and tasks seem to tap important teaching competencies which are needed by beginning teachers to teach effectively.
- For the most part, the teachers considered the content to be fair with respect to assessing diverse groups of teachers from varying teaching contexts.
- The majority of teachers who participated in the assessment judged the tasks to be job-related. In many cases, however, teachers indicated that they had not received instruction or training in performing these tasks.

- Knowledge of how to teach in a variety of contexts or to diverse student populations was not assessed well.

- The effect of experience in teaching focal topics on teacher performances, and the level of difficulty for new teachers are issues which could not be explored with available pilot data.

If the SSI-SM is selected for further development, it would benefit from a broad review by a panel of state and national experts in mathematics, mathematics pedagogy, performance assessments, and educational policy. California teachers and teacher educators should especially be a part of this panel. Such a review should be directed toward improving the representativeness of the content with respect to a secondary mathematics curriculum, clarifying the range of valid conclusions that can be drawn from performance on the SSI-SM, and identifying potential weaknesses in the instrument which could be remedied before incurring the expense of field testing.

Assessment Format

The semi-structured interview format will be discussed at length in Chapter 8 and contrasted with the classroom observation and multiple-choice examination methods of teacher assessment. Its strengths appear to be in the assessment of a teacher's ability to plan instruction and to understand the subject at a conceptual level of understanding. It appears weakest in assessing the ability to implement instruction and manage the classroom.

Summary

If semi-structured interviews are selected as a method of assessing new teachers for credentialing purposes, the SSI-SM has high potential for serving as a prototype not only for mathematics, but for other subjects. However, it needs further development of the scoring indicators, a closer alignment of the questions and the indicators, and pilot testing of the revised version before it can be considered ready for a field test.

CHAPTER 5:
SEMI-STRUCTURED INTERVIEW: ELEMENTARY MATHEMATICS

CHAPTER 5:

SEMI-STRUCTURED INTERVIEW: ELEMENTARY MATHEMATICS

The Semi-Structured Interview in Elementary Mathematics (SSI-EM) was developed by the Stanford Teacher Assessment Project (TAP). The version of the SSI-EM used in the Spring 1989 pilot test was not a final product, but rather a revised version of an assessment for certifying distinguished master teachers. Stanford had previously pilot tested an earlier version of the assessment with a sample consisting mostly of experienced teachers, but a few first-year teachers and student teachers had also been assessed. Based on this experience, the version used in the earlier pilot test was revised for use with beginning teachers.

The assessment consists of a series of semi-structured interviews addressing four tasks. The candidate performs a task and then is interviewed. The four tasks are described below.

(1) **Lesson Planning:** A teacher receives 30 minutes to plan a lesson on a given topic for a fifth-grade class, and then responds to questions about that lesson.

(2) **Topic Sequencing:** Using a set of 17 cards representing mathematical topics in a unit, a teacher sorts the cards into groups of topics, selects the cards representing the major themes of the unit, defines the topic on each card, and arranges eight of the cards in the order of perceived difficulty for students (least difficult to most difficult).

(3) **Instructional Vignettes:** A teacher responds to a series of hypothetical situations involving students in after-school tutoring sessions.

(4) **Short Cuts:** A teacher is presented with two purported computational shortcuts or rules of thumb for solving mathematical problems and evaluates them in terms of their pedagogical and mathematical soundness.

This assessment closely resembles the SSI-SM (which was the subject of Chapter 4). The SSI-EM and the SSI-SM are constructed in a similar manner in that they combine two assessment strategies, the semi-structured interview and the assessment center, which were described previously in the discussion of the SSI-SM. However, the set of tasks differ; those tasks which are most similar have slightly differing foci.

The administration of the assessment, the assessment content, and the assessment format are discussed below. The discussion of the SSI-EM concludes with a summary of our evaluations of the SSI-EM's potential as a prototype for further assessment development in elementary mathematics, as well as other areas of teacher performance.

Administration of Assessment

This section begins with an overview of the administration of the assessment. It is followed by a discussion of the logistics involved in arranging the pilot test.

Overview

The administration of the SSI-EM occurred between May 30 and June 24, 1989. As is seen in Table 5.1, a total of 41 teachers were interviewed, the majority of whom were female. Five minority teachers participated in the assessment. Grade levels reported by the teachers are also shown in Table 5.1. The table shows that while most of the teachers taught fifth or sixth grade, nearly one fourth taught a combination of grades. Two teachers taught in a middle school.

No data were collected on the length of the participants' prior teaching experience. Although all teachers participating in new teacher support projects were to be in their first or second year of teaching, a miscommunication resulted in the inclusion of some teachers with over five years of experience in the initial administration of the SSI-EM assessment.

Two different forms of the assessment were piloted. One focused exclusively on elementary fractions; the other consisted of two tasks (Lesson Planning and Topic Sequencing) that focused on ratio and proportions, and two (Instructional Vignettes and Short Cuts) that focused on elementary fractions. The number of teachers who participated in the pilot of each task is: 41 for Instructional Vignettes and Short Cuts; 25 for Lesson Planning (elementary fractions); 16 for Lesson Planning (ratio and proportions); 24 for Topic Sequencing (elementary fractions) and 17 for Topic Sequencing (ratio and proportions).

Logistics

Logistical activities for this assessment included (1) developing orientation information to be sent to teachers and principals, (2) identifying teacher samples, (3) identifying and training assessors, (4) scheduling the test administrations, (5) making facility and site arrangements, (6) gathering the materials to conduct the assessment, and (7) arranging payment to school districts and some teacher participants. Logistical activities are described in detail in the *Administration Report for Spring, 1989.*

As with the SSI-SM, orientation materials sent to the teachers were limited in scope. The tasks were identified and the structure of the assessment center activities was briefly described. Since the SSI-EM differs markedly from other assessments of teaching that are more familiar to teachers, the quality of the orientation materials affects the teacher's ability to anticipate activities and prepare for the assessment. The nature of these materials was previously described in the logistics section of the discussion of the SSI-SM.

TABLE 5.1

SEMI–STRUCTURED INTERVIEW IN ELEMENTARY MATHEMATICS: PILOT TEST PARTICIPANTS

(Total Number of Teachers = 41)

Descriptive Characteristics of Participants	Distributions of Participants
Grade Level	
4/5	4
5	16
5/6	4
6	14
6/7	2
Not Specified	1
Gender	
Male	10
Female	31
Ethnicity	
Asian	2
Black	2
Hispanic	1
White	34
No Response	2
Location of Assessment	
Anaheim	
Fresno	
San Diego	
Ventura	

Some difficulties in scheduling were experienced, which should be kept in mind if a semi-structured interview is considered for statewide adoption for credentialing purposes. Some teachers are on a year-round teaching schedule, complicating efforts to select convenient times for assessments. Those year-round teachers who began school in June have roughly three months more experience than teachers on more traditional schedules at the same point in the year. Our sense is that this difference in experience probably has little effect on performance on the SSI-EM; however, this should be investigated for all teaching competencies assessed.

As with the SSI-SM, locating appropriate facilities with large numbers of small rooms proved to be challenging. Scheduling assessment center activities at times of the year when schools and universities are not in session would make these facilities available and ameliorate this problem. However, there may be costs associated with reopening "closed" facilities (e.g., heat, custodial services).

All interviews were audiotaped with one tape recorder; the taping quality was checked at the beginning of each interview. For three tapes, data were lost due to failure to tape the interview at some point. This could be avoided by adopting a policy that once the tape recorder is started, it is never turned off, even if it records substantial time when a candidate does not speak as they are working on a solution to a problem. One additional interview was deemed inaudible by one scorer, but another scorer rated it with no reported difficulty.

Security

Slightly different versions of the assessment tasks had previously been administered to master teachers during the initial Stanford TAP pilot testing. Reports of the prototype testing which contained the previous protocols had been distributed by the Stanford TAP, though not widely. Therefore, only minimal security precautions were taken.

Assessors carried the interview protocols with them at all times or left them in securely locked rooms during assessments. Teacher notes were collected at the end of each task and disposed of at the end of each day.

As with all assessments, security of the documentation of teacher performance and evaluations is required. The extent to which a semi-structured interview needs redevelopment for each administration is unclear. The nature of a semi-structured interview is such that its security is compromised after each administration. While individual questions are not especially memorable, the tasks certainly are. One teacher who taught at the same school as a teacher who had taken the SSI-EM earlier in the week told us that the other teacher had described the experience. She believed that it was possible to learn a great deal about the test content from someone who had previously taken the test. If semi-structured interviews are to be used for credentialing and thus become a high-stakes assessment, the contents will be quickly made public.

Although teachers could ascertain the topics and the thrust of the questions, the degree to which prior preparation would substantially compromise the validity of the test is unclear. The Stanford TAP advocates informing teachers of the focal topic of the test in advance to allow teachers to prepare. The questions in each task are interrelated, making memorization of appropriate responses difficult. Some questions and scoring criteria assess a relatively sophisticated knowledge of mathematics and mathematics pedagogy which would be difficult to acquire in a short period of time; others would be more susceptible to memorizing formulaic answers.

The state of Georgia abandoned a semi-structured interview as a certification requirement when standardized answers were developed and taught to candidates, affecting the validity of the interview data. Whether the level of performance demanded by the SSI-EM is sufficiently complex to inhibit the utility of similar strategies is an issue which needs to be explored prior to its use for credentialing.

For security purposes, topics, and perhaps tasks, should be varied across assessment dates. However, examples of questions and responses which indicate the level of performance required to pass, together with descriptions of the tasks, should be given as information to candidates who are planning to participate in the assessment.

Assessors and Their Training

Assessors for the SSI-EM need to be knowledgeable about mathematics and mathematics pedagogy at the elementary level. They must also have good interviewing skills. FWL staff recruited seven California educators to administer the SSI-EM. All but two had taught elementary school. Two of these were retired administrators, and two others were elementary teachers on sabbaticals who had been working with a mathematics project at a local university. Of the two assessors with no elementary teaching experience, one had designed elementary math curricula, and the other had a strong math background. In addition to these, when necessary, three FWL staff members also served as assessors to complete an assessment team.

The seven recruited assessors, along with two staff members from FWL, received training in the administration of the SSI-EM. Training sessions were conducted on May 16 and 19 by one of the original test developers, a representative from the Stanford Teacher Assessment Project. Each assessor was trained to administer two of the tasks, either Lesson Planning and Topic Sequencing or Instructional Vignettes and Short Cuts. The training consisted of (1) an overview of the assessment project, including its purpose and its relation to the California New Teacher Project; (2) a general orientation to the purpose of the assessment; (3) an overview of each task; and (4) paired practice in administering the two tasks.

Assessors all felt that their training had been adequate, although two mentioned that they would have welcomed more discussion of the specific intent of the questions and skills being assessed. FWL staff who administered tasks felt that it was difficult to construct effective probes when it was unclear as to what information was actually being sought through the questions. This was less a problem in the design of training than a result of the stage of development of the scoring system, which was due to be extensively revised and therefore was not described to the assessors.

During this pilot test the assessors were allowed to ask probing questions at their own discretion. They were carefully instructed to use probes that aimed at clarification or expansion of a teacher's response or lack of response, and not to hint at a correct answer. Assessors were monitored during the training session to see if they were able to adhere to this admonition. However, feedback from scorers indicates that there were instances when teachers were guided to the correct answer.

Assessor opinions on how often assessors should be retrained if they did not administer the protocol at least monthly ranged from every three or four months to once a year. Two assessors thought that retraining was unnecessary, but either a review or an update would be useful if an assessor had not administered the task in some time or if changes had been made in the instrument.

Teacher and Assessor Impressions of Administration

Teachers and assessors were pleased with most aspects of the arrangements. When asked whether the arrangements (e.g., scheduling, facilities, distance to travel to assessment site, breaks and lunch, room arrangement) were reasonable, 30 of the 40 teachers returning surveys responded affirmatively, four teachers negatively, and two teachers responded positively but identified particular aspects that were unsuitable. Nine teachers commented that the assessment should have been scheduled earlier in the school year.

Scoring

During the previous administration of the SSI-EM tasks by the Stanford TAP, separate scoring systems for each task were developed. For the CNTP pilot test, a TAP representative developed a new scoring system with more similarity across tasks for scoring the SSI-EM. The scoring scale was also changed from a six-point scale to a two-point pass/fail scale, with "unable to score" being a third option. Specific components of the teachers' performance were rated. Ratings were generally holistic, though a few categories consisted of checklists of major points to be covered in the response. To pass, a teacher needed to pass more than two-thirds of the components, so passing rates were set at 70% for **Topic Sequencing** and **Instructional Vignettes**, while **Lesson Planning**, which had more components, required the teacher to pass at least 75% of the components. The scoring system for **Short Cuts** differed from the format of the other tasks in that the summative judgments for performance on each individual short cut and the entire task were holistic. Appendix B lists the scoring categories for each task.

The scoring system emphasized certain aspects of teacher knowledge, including:

- knowing various ways to organize topics in the discipline;
- knowing what's difficult, easy, and important within a topic and why;
- knowing multiple ways to represent topics that make it easy for others to comprehend; and
- anticipating misconceptions and preconceptions that students have about the content.

The interview protocols were revised on the basis of the previous pilot test experience to clarify the questions and eliminate questions which did not seem to elicit useful information. The scoring criteria were adapted as well. Criteria which were deemed to be too difficult for beginning teachers were dropped. The revised scoring criteria were also made more specific than the original criteria to resolve application problems that had been reported by scorers. Some portions of the interview were excluded from scoring because they did not elicit useful information.

Scoring Process

Scorers listened to taped interviews and rated the teacher responses with the use of task-specific categories. The number of categories per task varied from 14 for **Topic Sequencing** to 20 for **Lesson Planning**. There were no categories which were rated for more than one task. A few categories were check lists, e.g., the teacher mentioned three out of four specified aspects of the topic which students find difficult. Most required holistic judgments, e.g., the teacher's description of the strengths of a particular short cut was satisfactory. If sufficient information was not available to address a scoring category, the category was omitted for that teacher. To determine a pass/fail score, with the exception of **Short Cuts**, the number of categories receiving a passing score was divided by the total number of categories scored for each task and the percentage of categories passed was calculated. Passing scores were set at 75% for **Lesson Planning** and 70% for **Topic Sequencing** and **Instructional Vignettes** by the TAP representative. The reason that the passing scores differed between **Lesson Planning** and the other tasks was that there were fewer scoring categories in the other tasks, so each individual category had more weight when computing the final score. For **Short Cuts**, a teacher's evaluation of each of two algorithmic short cuts was rated separately. First, various aspects of the teacher's response were rated on a pass/fail dimension to assist in arriving at a pass/fail score for the entire set of responses for the short cut. The judgments for the two short cuts were compared. If they agreed, then the entire task received the common rating. If the two ratings differed, then the evidence was compared; if the weight of evidence did not clearly support an overall score (e.g., if the teacher received a marginal pass on one short cut and a marginal fail on the other), then the task was assigned the score of the first short cut, which was deemed to be more representative of a teacher's competence by the test developers.

Discussion of Scoring System

For the most part, the SSI-EM scoring system was not a check list of features in a teacher's response, but relied on professional judgments to evaluate teacher responses. This increased the ability of the system to apply across differing teaching styles and approaches, but it increased the difficulty in training scorers to the same standard, especially since examples representing the total range of potential responses were not available.

The scoring system does not generalize across tasks or even across the same task focusing on different topics. It is also insensitive to differences in the quality of responses; clearly superior responses receive the same credit as marginally acceptable ones. (This is not a problem for credentialing decisions, but limits its utility for professional development purposes.) Although scorers generally felt that the scores for each task reflected their summative evaluation of a teacher's performance, they also reported some occasions when they were unable to score what they considered to be significant aspects of a teacher's response due to the lack of a relevant scoring category. In its choices of scoring categories, the scoring system contains implicit judgments about the relative merit of particular aspects of a teacher's potential response. These choices are likely to be the subject of debate within the community of mathematics educators. Because of the greater degree of specificity compared to the scoring system of the SSI-SM, a professional consensus on scoring categories is likely to be more difficult to achieve.

As with the scoring system for the SSI-SM, the after-the-fact development of the SSI-EM scoring resulted in a misalignment between questions and scoring criteria. While assessors for the SSI-EM were invited to probe for clarification, they were unable to do so effectively because they did not know the scoring criteria. This resulted in numerous instances where there was insufficient information to rate some specific components of the responses.

Scorers and Their Training

FWL recruited scorers from the San Francisco Bay Area on the basis of referrals from local mathematics and science programs for teachers. The required qualifications for scorers were experience and training in mathematics education and/or required elementary education. Experience in conducting observations or evaluations of teachers was a desirable qualification. Due to unavailability or lack of proximity to the training site, only one of the seven assessors participated in the scoring. Of the eight scorers, five were present or former elementary school teachers. Three scorers were teachers on sabbatical; one was a professor of mathematics education at a local university; two were math education consultants; and two were doctoral students with interests in math and science education.

Each scorer was trained to score two of the tasks, and provisions were made for double scoring four to six tapes to serve as a rough reliability check. The training of scorers was conducted by the TAP representative and a FWL staff member; it consisted of two phases. In the first phase, each scorer received one-half day training per task. Scorers received the interview protocols in advance and became familiar with the protocols prior to the training. At the training session, scorers were given an overview of the California New Teacher Project and the SSI-EM. The scoring guide, which described each scoring category and how to rate it, was then handed out and explained in detail. Two transcripts of previously scored interviews were provided, and the reasons for the scoring were explained. Scorers then listened as a group to a tape and practiced scoring it. Then scores were compared and questions about application of the scoring rules were answered by the trainer. All scorers for a given task were then given the same four tapes to score as a preliminary reliability check. They returned a week later for the second phase of scoring training, meeting for an hour and a half with the trainer to compare their scoring decisions and clarify how to apply the scoring criteria.

At the end of the training, each scorer was given 18 or 19 tapes to score, including some which were to be scored by another scorer. They then scored the tapes and returned them to FWL.

This level of training did not prove to be adequate for most tasks. As a rough check on reliability, pairs of scorers rated four tapes for **Lesson Planning** and **Topic Sequencing** (however, one scorer found much of one tape for **Lesson Planning-Ratios** inaudible) and six tapes for **Instructional Vignettes** and **Short Cuts**. The percentage of agreements between raters on pass/fail judgments for each task was 50% for **Lesson Planning-Fractions**, 67% for **Lesson Planning-Ratios**, 100% for **Topic Sequencing-Fractions**, 25% for **Topic Sequencing-Ratios**, 100% for **Instructional Vignettes**, and 50% for **Short Cuts**. In many cases, the difference in rating particular scoring categories hinged on a single piece of evidence which one scorer had heard and the other had not. Sometimes, ratings varied due to differences in interpretation of teacher comments. This was especially common for responses deemed to be borderline.

All but one of the scorers rated their training as very good, with the other scorer rating the training as adequate. Suggestions for improvement included more examples spanning the range of performances, more opportunities to compare and discuss ratings with the other scorers, and reorganizing the scoring guide for **Instructional Vignettes** to correspond with the scoring sheet. Scorers generally believed that the scoring guidelines were complete and clear, but three scorers called for more detailed examples; they felt there was too little guidance for scoring answers which were on the borderline between acceptable and unacceptable.

Teacher Preferences about Feedback

Teachers were asked what types of feedback would be most useful and by whom, when, and in what format should the feedback be provided. Of the teachers expressing an opinion, the most popular response (by 18 of the 40 teachers) was that feedback should identify the strengths and weaknesses of a teacher. Eight teachers desired suggestions for improvement, three wanted a summary of all results and scores, three wished to know if they had given the correct answers, and three wanted the feedback to explain the purpose of the assessment and to identify evaluation criteria.

Teachers also gave a range of answers about other questions related to feedback. Of those specifying a time frame, seven said immediately (presumably by the interviewers), and three said as soon as possible after the assessment. In terms of format, seven teachers preferred a written format, three said oral (in addition to the seven who said immediately, which presumably would be oral), and three specified either written or oral feedback. One teacher requested that the feedback be mailed to a teacher's home. In terms of who should provide the feedback, nine teachers requested an impartial party such as a testing service; seven selected the interviewers; four specified another educator (such as a mentor teacher); and two felt that feedback should be channeled through the New Teacher Support Project.

Assessment Content

The SSI-EM was originally designed to assess master or expert teachers. As a result, the tasks emphasize both mathematical knowledge and state-of-the-art knowledge of how to teach elementary mathematics (pedagogical content knowledge). More general pedagogical skills, such as classroom management, receive little attention. Thus, the

SSI-EM is aimed at measuring subject-matter competency, and not more general teaching competencies.

This assessment was neither developed for beginning teachers nor to be congruent with California's curricular emphasis. However, an analysis of the appropriateness of its content for beginning teachers and its congruence with the emphases of the State mathematics curriculum guides and frameworks and California Standards for Beginning Teachers can suggest the form that an assessment of beginning teachers in elementary mathematics could take.

This section evaluates content-related aspects of the SSI-EM along the following dimensions:

- Congruence with Curriculum Guide and Framework emphases;
- Coverage of the California Standards for Beginning Teachers;
- Job-relatedness;
- Appropriateness for beginning teachers;
- Appropriateness across teaching contexts;
- Fairness across groups of teachers; and
- Appropriateness as a method of assessment.

Congruence with emphases of the relevant curriculum guide is addressed first.

Congruence with California Curriculum Guides and Frameworks

The 1985 *Mathematics Framework for California Public Schools: Kindergarten through Grade Twelve* identifies five major areas of emphasis: (1) problem solving, (2) calculator technology, (3) computational skills, (4) estimation and mental arithmetic, and (5) computers in mathematics education. In the SSI-EM, one or more tasks addressed the areas of calculator technology, computational skills and estimation. Problem solving was indirectly addressed, while mental arithmetic and computers in mathematics education were not addressed at all. The specific ways in which the areas were included in the assessment are described as follows.

- Although problem solving was not a direct area of focus of the assessment, the SSI-EM scoring criteria stress conceptual understanding of mathematical algorithms and terms, which facilitates the application of algorithms and concepts to new problems. Teachers also sometimes needed to provide more than one approach to a mathematics problem to receive credit for their solution, which reflects the multiple strategies approach to teaching problem solving that is stressed in the *Mathematics Framework.* Two of the four vignettes in **Instructional Vignettes** could be treated as problem solving situations by the teachers, but they were not required to do so.

- o All four situations in **Instructional Vignettes** involved a student's use of a calculator and resulting misconceptions.
- o The evaluation of the teaching of computational skills by the SSI-EM is consistent with the *Mathematics Framework*'s emphasis on conceptual understanding of why algorithms work.
- o One situation in **Instructional Vignettes** addresses estimation errors and their remediation.

The *Mathematics Framework* also emphasizes the following characteristics in terms of delivery of instruction in mathematics: teaching for understanding, reinforcement of concepts and skills, problem solving, situational lessons, use of concrete materials, flexibility of instruction, corrective instruction/remediation, cooperative learning groups, mathematical language, and questioning and responding. The main theme of the SSI-EM is teaching for understanding; every exercise has multiple scoring criteria which address this teaching characteristic. A major focus of **Short Cuts** is whether the teacher is promoting computational efficiency at the expense of conceptual skills. With the exception of cooperative learning groups, the remaining instructional characteristics are embedded in one or more tasks. Some tasks, however, could be modified slightly to strengthen measurement of the appropriate use of these techniques. **Lesson Planning**, for example, could be modified to address situational lessons by asking the teacher to explain why they either did or did not include this approach in the lesson. (The scoring would address the ***appropriateness*** of either use of a situation or the ***rationale*** for not using such an approach, not the use or nonuse of a situation.)

The congruency of the SSI-EM with the emphases in the *Mathematics Framework* is summarized in Table 5.2.

Extent of Coverage of California Standards for Beginning Teachers

The California Beginning Teacher Standards are criteria for teacher competence and performance which the Commission on Teacher Credentialing expects graduates of California teacher preparation programs to meet. Listed below are brief descriptions of Standards 22 through 32 (with each standard following in italics). To evaluate this assessment instrument and make inferences about the assessment approach which it represents in terms of the appropriateness for use with California elementary mathematics teachers, the SSI-EM tasks and scoring criteria were compared with the 11 California Beginning Teacher Standards. Although some of the questions in the SSI-EM task elicited information pertaining to a particular standard, the scoring criteria often failed to capitalize on this information. This will be noted in the discussion of the standards where it occurs. Each standard will be discussed separately.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* None of the content in the SSI-EM addresses this standard. Indeed, except for the "clearly stated expectations regarding student conduct" (CTC, 1988: 23), the other

TABLE 5.2

COVERAGE OF THE CALIFORNIA MATHEMATICS FRAMEWORK BY THE SSI-EM

Content	Method of Coverage	Extent of Coverage
Areas of Emphasis:		
Problem Solving	-In general, many scoring criteria address prerequisites for problem solving. Instructional Vignettes focus on remediating student errors in solving problems.	Partial
Calculator Technology	-Instructional Vignettes focuses on student errors resulting from the use of a calculator.	Full
Computational Skills	-Tasks and scoring criteria emphasize underlying a base of conceptual understanding for developing computational skills.	Full
Estimation and Mental Arithmetic	-One problem in Instrutional Vignettes focuses on estimation errors; mental arithmetic not addressed.	Partial
Computers in Mathematics Education	-Not addressed.	None
Delivery of Instruction:		
Teaching for Understanding	-Implicit in all tasks and scoring criteria.	Full
Reinforcement of Concepts and Skills	-One scoring criterion of Lesson Planning task.	Partial
Problem Solving .	-In general, many scoring criteria address prerequisites for problem solving. Instructional Vignettes focus on remediating student errors in problem solving.	Partial
Situational Lessons	-Some vignettes address situations.	Partial
Use of Concrete Materials	-Appropriate use addressed by two scoring criteria for Lesson Planning.	Partial
Flexibility of Instruction	-Scoring criteria for Lesson Planning and Shortcuts address multiple methods of presenting concepts or solving problems.	Partial
Corrective Instructions Remediation	-Instructional Vignettes and Lesson Planning address remedial instruction.	Full
Cooperative Learning Groups	-Not addressed.	None
Mathematical Language	-Addressed by scoring criteria for each task.	Full
Questioning and Responding	-Focus of Instructional Vignettes.	Partial

factors such as rapport with students address teacher behavior when interacting with students. This would be difficult to simulate in an interview situation, except through vignettes or videotapes.

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies, activities, materials and assessment plans that are well defined and coordinated with each other.* This is addressed in depth by two tasks in the SSI-EM: **Lesson Planning** and **Topic Sequencing.** In **Lesson Planning**, the teacher is asked to plan a three lesson sequence, with the middle lesson described at length. Three of the scoring criteria, counting for approximately 15% of the total score, address the coordination of the lessons and the adequacy of the amount of practice devoted to two concepts taught in the lessons. **Topic Sequencing** requires the teacher to group mathematical topics according to how they should be taught. The scoring criteria do not address the appropriateness of the grouping, focusing instead on measurement of the teacher's understanding of the selected mathematical topics and their interrelationship. This content knowledge is necessary to plan effective instruction.

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities and materials that are appropriate for students with diverse needs, interests and learning styles.* This standard is addressed to some extent by **Lesson Planning** and **Instructional Vignettes**, though not in depth. In **Lesson Planning**, two of the scoring criteria, representing 10% of the total score, address the use of multiple representations of the content in presentations or responses to student questions; three of the scoring criteria, constituting 15% of the score, address prior student knowledge necessary to understand the concepts being taught. One series of questions asking what would cause deviation from the plan and how a teacher monitors student understanding was not scored. The teacher's responses to this section of the protocol would yield information about competencies addressed by this standard. **Instructional Vignettes** has four scoring criteria constituting 20% of the total score which evaluate the appropriateness of the teacher's understanding of student thinking in vignettes that portray student misconceptions or confusions. However, the addition of questions and/or probes addressing assumptions about the sources of student errors would facilitate scoring these criteria, which were often left unscored because of the lack of information to judge the appropriateness of the teacher's response. To fully address this standard would either require development of a new task or substantial revision of **Lesson Planning** or **Instructional Vignettes.**

Standard 25: Student Motivation, Involvement and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* Information about motivation and the involvement of students in the development of the lesson is elicited by **Lesson Planning**, though not in great depth. **Lesson Planning** contains two scoring criteria constituting approximately 10% of the total score that address motivation and involvement of students in the lesson. The appropriate use of reinforcement and feedback, setting high standards, equitable treatment of students, and discipline are not addressed by the SSI-EM, and would require the development of additional questions and/or a new task.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* This standard addresses the linguistic complexity of a teacher's communications; three of the four tasks in the SSI-EM address both this *and* conceptual aspects of a teacher's communication with students. **Lesson Planning** has one criterion addressing the conceptual clarity of the introduction to the lesson and two criteria that address the appropriateness of the teacher's response to a student error, accounting for approximately 15% of the total score. **Topic Sequencing** has three criteria directly addressing either the language used to explain concepts or the knowledge of common conceptual understandings of students. These criteria constitute 21% of the total score. **Instructional Vignettes** has four criteria constituting approximately 20% of the total score which address the adequacy of teacher explanations, both in terms of the clarity of the communication and in terms of laying a foundation for mathematical concepts introduced later in the curriculum. It is not clear whether or not an interview would capture nonverbal communication by a teacher; several teachers mentioned that it was difficult to respond as if the interviewer were a student.

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* One section in **Lesson Planning**, which was not scored, addressed the routine monitoring of levels of student achievement during the lesson. Four scoring criteria in **Instructional Vignettes**, constituting approximately 20% of the total score, addressed the identification of student errors. Skills in constructing and interpreting summative forms of evaluation are not addressed.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* Student outcomes are not directly addressed by the SSI-EM. However, many of the scoring criteria focus on whether the teacher is laying a cognitive foundation that enables the student to achieve understanding of mathematical concepts and their interrelationship. For example, one of the criteria for scoring **Topic Sequencing** is whether the metaphors and analogies, if any, used to explain concepts facilitate or hinder understanding of the concepts.

Standard 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward the subjects learned, the students themselves, and their capacity to become independent learners.* The encouragement of positive interaction among students and the provision for independent learning experiences are not addressed by any task in the SSI-EM. Student motivation was discussed under Standard 25.

Standard 30: Capacity to Teach Cross-culturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural, gender, linguistic and socioeconomic differences.* This standard is not addressed by the SSI-EM; to do so would require adaptation of the tasks or development of new tasks. This is further discussed in a subsection on the appropriateness of the SSI-EM for assessing teachers who teach diverse students.

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* This standard refers to student teaching experience, although it could be interpreted to apply to the ability of teachers to accept teaching assignments that span the elementary grades. The SSI-EM concentrates on a single grade level; it could be constructed, however, so every task would address a different topic at a different grade level. If this approach were taken, some of the scoring criteria in the SSI-EM, such as identifying what students find difficult about a topic, would become problematic because teachers who have taught the topic would be advantaged relative to those who have not.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and develops professionally through self-assessment and collegial interactions with other members of the profession.* The SSI-EM does not address this standard. Since the SSI-EM focuses on content pedagogy, any task constructed to measure this standard would be qualitatively different from the other tasks, all of which focus on the teaching of elementary mathematics.

The extent to which the SSI-EM covers the California Standards for Beginning Teachers is summarized in Table 5.3.

Job-relatedness

Teachers, assessors, and scorers were asked their opinion of the assessment's job-relatedness. Thirty-one of the 40 teachers agreed that "all the major tasks (i.e., Lesson Planning, Topic Sequencing, Instructional Vignettes, and Short Cuts) composing this assessment are relevant to their job of teaching"; eight felt they were not. Several teachers commented that their students asked them some of the same questions contained in **Instructional Vignettes** and **Short Cuts**. Five teachers singled out **Instructional Vignettes** as being irrelevant, and four each identified **Topic Sequencing** and **Short Cuts** as irrelevant. One of these teachers identified both **Instructional Vignettes** and **Short Cuts** as irrelevant.

The teachers expressed a variety of reasons for judging some of the SSI-EM content to be irrelevant. One teacher was not sure that topics need to be taught in sequence; another observed that texts sequence topics; another objected to the focus on methods and mathematical reasoning in **Short Cuts**; and another felt that calculators received too much focus in **Instructional Vignettes** compared to their representation in the curriculum. More than 75% of the new teachers considered the SSI-EM to be job-relevant, however.

Most of the assessors and scorers tended to feel that the SSI-EM tasks reflected a teacher's responsibilities in the classroom. The one task that some assessors did not feel was related to a new teacher's experiences was **Topic Sequencing**, since this is prescribed by textbooks. However, assessors felt that **Topic Sequencing** reflected a perspective on instructional design that would be desirable for a teacher to have. Scorers of **Topic Sequencing** believed that elementary teachers, and especially fifth grade teachers, need to be able to sequence topics to effectively plan instruction.

Appropriateness for Beginning Teachers

The degree of appropriateness for beginning teachers was judged from two kinds of evidence: (1) the perceptions of teachers, assessors and scorers, and (2) the performance of the teachers on the assessment tasks.

TABLE 5.3

EXTENT OF COVERAGE BY THE SSI–EM OF CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Standard	Method of Coverage	Extent of Coverage
22: Student Rapport and Classroom Environment	–Not covered.	None
23: Curricular and Instructional Planning Skills	–Addressed in depth by Lesson Planning and Topic Sequencing.	Full
24: Diverse and Appropriate Teaching	–Partially addressed by Lesson Planning and Instructional Vignettes, though not in depth.	Partial
25: Student Motivation, Involvement, and Conduct	–Ability to motivate and involve students assessed in Lesson Planning and scored by two criteria.	Partial
26: Presentation Skills	–Scoring criteria for three tasks that address conceptual clarity and appropriateness of teacher explanations.	Partial
27: Student Diagnosis, Achievement, and Evaluation	–Focus of one task.	Partial
28: Cognitive Outcomes of Teaching	–Not directly covered.	None
29: Affective Outcomes of Teaching	–Not covered.	None
30: Capacity to Teach Crossculturally	–Not covered.	None
31: Readiness for Diverse Responsibilities	–Not covered.	None
32: Professional Obligations	–Not covered.	None

Perceptions. For the most part, teachers felt that, as new teachers, they had "sufficient opportunity to acquire the knowledge and abilities needed to respond in a reasonable manner to the assessment questions and tasks," with 28 of the 40 teachers marking "yes," nine marking "no," and one giving a qualified answer. However, 15 of the teachers felt the **Topic Sequencing** task was too difficult, and three teachers each identified **Instructional Vignettes** and **Short Cuts** as too difficult. Seven other teachers specifically criticized the content of the assessment as being too difficult, identifying different aspects such as the lack of supplementary materials or questions which asked them to explain why fractions are useful in real life or how to teach material they had not previously taught. As will be seen in the discussion of the teachers' performance, teachers generally exhibited gaps in content knowledge which might make it difficult to evaluate supplementary materials or identify applications of mathematical concepts. Describing how to teach unfamiliar material is difficult for beginning teachers, but the focal topics of the SSI-EM were all part of the elementary curriculum covered by the multiple subjects credential. Including topics previously taught would afford teachers maximal opportunity to draw upon their teaching experience as well as their knowledge of mathematics pedagogy. However, since the credential covers a broad range of grade levels, it is inevitable that an assessment consisting of an adequate sample of grade levels covered by the credential would include some topics which a beginning teacher had not taught.

Two teachers suggested that assessment be delayed until the second year, echoing comments by second-year teachers that they were glad they had one full year of experience prior to the assessment because they would not have done as well had they been administered the assessment in their first year. Such feelings are articulated well in the following comment:

> *I think it would be extremely difficult for a beginning teacher to demonstrate a competent understanding. The first year should entail* **in-services** *or further professional instruction by peers, etc. Assessment should be during the second year.*

Another teacher felt that questions about curriculum (which were not scored) were unfair:

> *New teachers are not generally aware of the abilities or curriculum for any particular grade level. They learn this after they are hired.*

Despite teachers' perceptions of adequate preparation, many teachers had difficulty describing mathematical concepts. Teachers were often at a loss when asked to provide a mathematical justification for a solution to a problem. This was true even for some teachers who had just correctly explained how to work the problem. As one assessor commented:

> *Instructional Vignettes seemed about right in content. However, while many teachers could describe the steps they would take in teaching or solving the problem, they had trouble naming the concepts.*

The two assessors who were classroom teachers felt that the assessment was too focused on mathematical sophistication and on content that was relatively difficult for beginning teachers. For teachers who were struggling with the content, it was difficult to display their skills in pedagogy and pedagogical content knowledge. Most teacher preparation programs do not require extensive courses in math methods; often a single course covering grades K-8 is all that is required. People who choose elementary teaching as a career are not required to have an extensive background in mathematics, and they may take several years to be comfortable with content that is included in the fifth- or sixth-grade mathematics curriculum. However, if the elementary mathematics curriculum is to be upgraded in line with the expectations of the *Mathematics Framework*, teachers will need a more sophisticated understanding of mathematics very similar to that required by the SSI-EM.

Scorers generally thought that the SSI-EM is a good prototype for assessing new teachers in the area of elementary mathematics, but identified shortcomings in specific areas where they believed that too much was expected of new teachers. These areas included: (1) the complexity of the evaluation of one of the Short Cuts (which did not work for a specific group of numbers) within the limited time period provided; (2) asking beginning teachers to depart from textbook orderings of lessons, which requires a great deal of professional self-assurance; (3) seeing the limitations of "short cut" algorithms, which depends on familiarity with student error patterns; and (4) ranking a set of topics in terms of student difficulty when students find most of the topics difficult.

Performance on assessment tasks. Performances on the specific tasks indicate that the majority of the teachers were ill prepared to adequately respond to the questions in the SSI-EM. While this was partly due to the original focus on the identification of exemplary master teachers, it was also due partly to weak content knowledge.

Table 5.4 shows the performance of the teachers on the tasks. Not all teachers are included in the table. Five of the 164 tapes could not be scored due to a failure to record the entire interview.

Although teachers felt most comfortable with **Lesson Planning**, they did not tend to do well on it. This may have been because the teacher's plan was evaluated on the basis of its capacity to foster conceptual understandings among students and not according to characteristics of its format. Criteria which most teachers were unable to meet included:

- communicating to students when the process of factoring was complete (i.e., how to know when you have found the answer);
- providing an adequate amount of practice for factoring (a concept which students find very difficult, but which is critical for that particular lesson); and
- explaining why there can be percentages greater than 100.

Although some of the teachers presented good application problems at the beginning of the lesson to capture student interest, almost half of the lessons on fractions were judged to be inadequate in motivating students. Most teachers ***did*** do well on:

TABLE 5.4

SEMI-STRUCTURED INTERVIEW IN ELEMENTARY MATHEMATICS: SCORING RESULTS

Task	No. of Teachers*	No. Passing	No. Failing	% Passing
Lesson Planning Fractions	25	12	13	48%
Lesson Planning Ratios	15	7	8	47%
Topic Sequencing Fractions	23	5	18	22%
Topic Sequencing Ratios	17	7	10	41%
Instructional Vignettes	38	24	14	63%
Short Cuts	40	22	18	55%

	Number of Tasks Passed*				
	0	1	2	3	4
Number of Teachers	4	9	12	6	5

*Due to five failures to record the entire interview, the number of teachers does not always total to 41. The number of tapes affected differed by task.

- actively involving the students in the lesson (e.g., asking students questions during the lesson);
- addressing both procedures and conceptual understandings during the lesson;
- keeping a smooth instructional flow between the sequence of three lessons described;
- providing a clear introduction to the lesson; and
- for the lesson on fractions, providing more than one representation of the content.

Teachers also had great difficulty in explaining the topics in **Topic Sequencing**. If definition by example had not been allowed, the scores would have been even lower. While a few of the topics were not clear from the title on the cards (e.g., fractions as a region/set), all of the topics were basic concepts or applications from the fifth grade curriculum. Some of the scoring categories did not seem to both scorers and FWL staff to follow closely from the questions asked in the interview. Additional probing or asking a direct question would have given the teachers a clearer idea of the type of response desired.

Most teachers scored poorly because they did not do the following in their discussion of topics:

- perceive the importance of cross-multiplication in a unit on fractions;
- perceive the importance of finding the percent of a number in a unit on ratios and proportions;
- provide an explanation of why common denominators are needed to add and subtract fractions; and
- defend their choice to add or delete topics.

In many instances, a concept which is clearly antecedent to another was placed well after it in the teacher's ordering. The teachers did best at identifying one or two of the most difficult topics to teach and explaining why students found them difficult.

The teachers performed best on **Instructional Vignettes**, which required them to provide remedial instruction to students making errors in solving problems with the use of calculators. This task was the only one in which any of the teachers received perfect scores, with six teachers doing so. Teachers generally could identify the student error and appropriately discuss the error with the student. Some teachers' spontaneous explanations were not only conceptually appropriate and creative but extremely clear and concise, using metaphors, examples or analogies which would appeal to fifth-grade students. One example is the teacher who explained why students need to learn "long ways" to compare the value of two fractions instead of just using a calculator:

> *You always need to learn the long way before you learn short cuts. It's like you learn the way to school and once you know the way to school, then you can cut across that dirt path, but if you cut across that path first, then you're not going to know where you're going to end up, so you* **have** *to know the long way and then you can make up your own short cuts.*

Over one-third of the teachers, however, could not figure out how to convert a recipe for six people to a recipe for eight people, as required by one of the vignettes, and as a consequence were unable to explain the problem to their students.

Teachers also had difficulty with key elements of **Short Cuts**, which required them to evaluate two algorithms for simplifying either the reduction or comparison of fractions. In discussing each short cut, teachers had difficulty with the following:

- identifying limitations;
- justifying whether or not they would teach it;
- describing ways to facilitate teaching it;
- providing a mathematical rationale for why it does or does not work; and
- identifying whether or not it works for all fractions.

One short cut was criticized by the teachers, scorers, and FWL staff for the complexity of the reasoning required to figure out whether or not it works. It was too complex to comprehend in a short period of time. However, teachers also had difficulty with the other short cut, which *did* work because of the identity principle. Teachers should be familiar with this concept, if not its name, when they teach math to intermediate students.

If the beginning teachers who participated in the SSI-EM are typical of recent graduates from teacher preparation programs (and there is no apparent reason to believe that they are not), then it would seem that beginning teachers are not equipped with an understanding of either mathematics or mathematics pedagogy sufficient to perform well on an assessment modeled after the SSI-EM. The low levels of performance are probably due partially to the original focus of SSI-EM on assessing master teachers. However, many of the assessors, scorers and FWL staff felt that the level of content knowledge exhibited by many teachers compromises their ability to teach mathematics to elementary students. Other work has found that teachers who can solve mathematics problems do not necessarily have an understanding of the underlying mathematical concepts and relationships (Leinhardt and Smith, 1985). This knowledge is needed for competent design of instruction.

It is often said that teachers, especially elementary teachers, do not need to be specialists in mathematics. Research on teaching of elementary mathematics finds that the quality of the developmental portion of a lesson differs considerably between effective and ineffective teachers (Good and Grouws, 1975). This is defined as:

> *The developmental portion of a mathematics period is that part of a lesson devoted to increasing comprehension of skills, concepts, and other facets of the mathematics curriculum. For example, in the area of skill development, instruction focused on why an algorithm works, how certain skills are interrelated, what properties are characteristic of a given skill, and means of estimating correct answers should be considered part of developmental work. In the area of concept development, developmental activities would include initial instruction designed to help children distinguish the given concept from other concepts. Also included would be the associating of a label with a given concept. Attempts to extend ideas and facilitate transfer of ideas are a part of developmental work (p. 114).*

It is precisely these skills which most of the participating teachers lacked. Furthermore, these skills, which go considerably beyond the ability to work the problems in the textbook, do not necessarily develop fully with experience. The *Mathematics Framework* takes the position that current teachers of elementary mathematics need an understanding of mathematical concepts and their interrelationships in order to make effective instructional decisions. While there are many instructional strategies that enable most students to work the problems in the textbook, some choices are superior to others in facilitating both mathematics instruction later in the curriculum and mathematical applications in daily life. The *Mathematics Framework* acknowledges that this goal requires a more rigorous preparation in mathematics education than most elementary teachers currently receive. It is quite likely that achieving the curriculum goals in the *Mathematics Framework* will require performances on the level of the SSI-EM.

Regardless of whether or not an assessment such as the SSI-EM is adopted, the performances of the teachers add credence to suggestions (e.g., Lampert, 1988) that teacher preparation programs need to strengthen their instruction in mathematics and mathematics pedagogy. Such strengthening cannot occur by requiring additional courses which concentrate on problem-solving algorithms. Instead, the additional preparation must focus on teaching the concepts and principles that underlie problem-solving algorithms. One California university has developed a four-week summer workshop which teachers attend to gain the skills necessary to implement the *Mathematics Framework*. Teachers increase their knowledge of elementary mathematics topics, and learn problem solving and group activities, which helps to equip them to implement the *Mathematics Framework*. However, the same institution's mathematics methods course for elementary teachers has not incorporated similar instruction because of time constraints within the current course. The course cannot be lengthened due to competing priorities within the year-long multiple subject credential program.

Comparison of beginning and experienced teachers. The four tasks comprising the SSI-EM were part of a larger set of tasks that was initially administered to teachers with varying amounts of experience by the Stanford TAP. In this initial administration, six teachers had more than 10 years of teaching experience, while seven had two years or less, allowing comparisons between beginning and experienced teachers. All tasks were scored on three dimensions: Command of Subject Matter, Content-Specific Pedagogy, and Pedagogical Sensitivity and Responsiveness to Students. New and experienced teachers differed least on the Command of Subject Matter dimension, which measured knowledge of mathematics as a discipline, i.e., its structure, boundaries and substance. Weak differences between the two groups were found for the Content-Specific Pedagogy dimension, which evaluated the ability to present mathematical knowledge in a way that facilitates student learning. Strong differences were found between more and less experienced teachers on the Pedagogical Sensitivity and Responsiveness to Students dimension. Here the assessors examined descriptions of teacher-student interactions, including engaging students, providing appropriate feedback, and establishing interpersonal relationships with students. Instances in which scorers felt there was not enough information to reliably rate a particular criterion were markedly more frequent for novices than for experienced teachers, particularly within the dimension of Pedagogical Sensitivity and Responsiveness to Students.

Appropriateness across Contexts

FWL also evaluated the appropriateness of each assessment across varying teaching contexts. Twenty of the 40 teachers providing comments on the SSI-EM did not feel it was a good measure of teaching ability across grade levels and subject areas and across different student groups and in different school/community settings. (Ten teachers felt it was; four teachers gave a qualified "yes" answer; and six had no response.) Most of the criticisms related to grade-level/subject-matter differences and diverse students.

Grade level and subject matter. Regarding the appropriateness of the SSI-EM for use in credentialing, several teachers were concerned about the inexperience of new teachers with the specific content of this instrument. Although this assessment was designed for and administered to fifth- and sixth-grade teachers, some of the teachers commented that they had not yet taught the material on which they were being assessed. Commented one teacher, "A new teacher is best able to talk about that area on which she spends time." Another teacher remarked, "Fractions is a hard concept to teach or think about how you'd teach it if you haven't already tried." One scorer who was a math education professor believed that some competencies, e.g., the ability to identify aspects of a lesson that were difficult for students, depended strongly on experience teaching that lesson. This theme of experience extended into other areas as well. Teachers in one district had used the same textbook that was used in the assessment; several of these teachers indicated that their familiarity with the text and the way it was organized helped them in the **Lesson Planning** task.

Teachers who had taught the topic were undoubtedly advantaged in drawing upon their experience to answer *some* questions, such as areas of student difficulty and activities that motivated students. However, these areas did not constitute the majority of the scoring criteria. Teachers could pass each task in the assessment if they could clearly explain mathematical concepts which are fundamental parts of the upper-elementary school curriculum, and determine which skills were necessary to learn specific mathematical concepts. This ability does not depend on experience teaching the topic, but upon familiarity with the topic. All teachers with multiple subject credentials are likely to teach these concepts, so measuring their level of content and pedagogical knowledge for topics which they have not taught does not seem to be unreasonable.

The influence of experience in teaching particular concepts is an issue that extends beyond the grade level at which the concepts are ordinarily taught. There was much concern among the assessors about the suitability of this assessment for primary-grade teachers, and concurrent interest in whether there would be two versions of the assessment for primary and upper-elementary teachers. The SSI-EM has a narrow focus, and any future assessment for the multiple subject credential would need to be more balanced in representing concepts across the whole elementary mathematics curriculum. To design an assessment in a single subject area, such as mathematics, which is suitable for teachers at all levels represented by the multiple subject teaching credential is challenging. However, as the name of the credential signifies, elementary teachers teach many subjects, which further compounds the difficulty of assessment design.

Diverse students. Some teachers felt that the assessment did not take into account teaching differing student populations. One teacher, for example, expressed the belief that an assessment would not be appropriate if the questions were not geared specifically to the types of students (e.g., low-ability, LEP) that a teacher has been teaching. Assessors reported that teachers who had taught low ability students (who were consequently at earlier points in the mathematics curriculum than the focal topics of the SSI-EM) seemed to have difficulty in drawing on their experience in answering the interview questions.

The SSI-EM needs to be improved in of assessing a teacher's ability to work with diverse students, either heterogeneous classes or specialized student populations. Since these types of classrooms are increasing, the revision of the SSI-EM to address this issue is not just an issue of fairness to teachers in differing contexts. It is also a matter of including all important teaching skills. Within a semi-structured interview format, one way to address this issue would be to construct vignettes describing children or classrooms with particular characteristics, and ask teachers to describe how they would construct a particular activity in the specified context. The vignettes would need to be carefully constructed to avoid stereotyping particular groups.

Fairness across Groups of Teachers

Over half of the teachers felt that the assessment was fair to new teachers of both genders, different ethnic groups, different language groups, and other groups of new teachers (26 of the 40 teachers agreed, nine disagreed, one gave a qualified answer). Of those who disagreed, the only reason given by more than one respondent was suggested by four teachers who felt that teachers from different linguistic groups would be disadvantaged because of the verbal skills required by the interview format.

None of the participating teachers was limited in English proficiency, though there was at least one teacher of a bilingual classroom. Not surprisingly, then, none of the assessors or scorers mentioned fairness to teachers of differing linguistic ability as a concern. However, there was some informal discussion of whether highly verbal teachers have an advantage over less verbal teachers.

The concerns related to fairness for different groups of teachers that were mentioned by the assessors on their feedback forms were: fairness across age groups, fairness to minority teachers, and fairness to teachers for whom mathematics was not a strength. During informal discussions, one group of assessors concluded that both very young and very old teachers seemed to have difficulty with the assessment. The young teachers seemed to be especially nervous and anxious about their performance. The older teachers usually were coming to teaching after a long absence from formal schooling, and seemed to have special difficulties with the terminology of the questions. The number of teachers at the extremes was very small, and their individual scores varied considerably, so no firm conclusions can be drawn.

In terms of the number of SSI-EM tasks that were passed, the six minority teacher were not statistically different from the 34 non-minority teachers, but the samples sizes were too small to warrant making firm conclusions. Assessors at one site reported concern about the performance of one minority teacher in particular, which raised questions in their minds about the assessment. They felt that this particular teacher showed great commitment to and potential for motivating inner-city students. However, the teacher's performance on the assessment demonstrated particularly weak content knowledge. This teacher expressed concerns about his/her inadequate content knowledge, acknowledging that the assessment was fair, since students often asked the same kinds of questions.

The assessors felt that this teacher showed promise as a teacher, but needed more strength in math content knowledge. They also felt that this teacher was likely to seek out assistance and benefit from it if it were offered. They were concerned about whether a state assessment for certification would allow or provide needed support. This is particularly important in view of the difficulty of attracting teachers to work in the inner city, the teaching profession's difficulty in attracting minority teachers, and the percentage of minorities failing current methods of assessment.

The assessors also raised questions about the fairness of the SSI-EM among teachers for whom mathematics was not a strength. They pointed out that elementary teachers must attain competence in a number of subjects requiring different skills. It may not be reasonable to expect teachers to attain the same degree of proficiency in all subjects. Although the state has expressed an interest in designing an assessment system which allows for particularly strong performances in one area to compensate for weaknesses in another, it is a policy decision whether "area" should be extended to apply to subjects as well as to specific teaching competencies.

Since the SSI-EM covered one of the most difficult mathematical topics that are taught at the elementary level, assessors also questioned whether the level of performance exhibited on these tasks was representative of a teacher's ability to teach less pedagogically difficult topics. This would be particularly important for those teachers who are less familiar with mathematics.

Appropriateness as a Method of Assessment

While teachers believed the SSI-EM tasks were fair, they were almost evenly split on whether the subject matter and concepts were appropriate for demonstrating their teaching skills, with 16 of the 40 teachers marking "yes," 17 marking "no,", and three giving qualified answers. Of the 17 teachers who marked "no," six explained that they had not previously taught the topics, and felt that their performance was not representative of their teaching skills. Seven participants objected to the narrow focus on one topic. Six teachers did not feel that the assessment conditions were realistic for various reasons such as the pressurized context and the limited extent of preparation.

As discussed earlier, we share the teachers' reservations about the limited focus. The decision whether or not to assess teachers on topics which they have not taught depends on a policy decision concerning the information to be gained from pilot testing. On the one hand, teachers are credentialed to teach across grade levels and topics. On the other hand, if competencies are being assessed which are assumed to depend upon experience, then it would be most appropriate to assess teachers in areas in which they had experience teaching. The teachers' reservations about the "realism" of the assessment conditions should be evaluated in a similar light. If the intent is to see a teacher's pedagogical decisions in the best light, then more time should be provided for planning a lesson, and supplementary materials should be available. On the other hand, it seems unlikely that any of the improvements suggested by the teachers would result in anything but marginal differences in the display of pedagogical content knowledge. If teachers do not understand basic mathematical concepts and their interrelation, their choices are not likely to improve given either additional time or supplementary materials whose quality they are unable to evaluate.

Reflecting their limited knowledge of tasks which they had not administered, assessors and scorers tended to be task specific in their perceptions of the ability of the assessment to measure teaching competency. In informal discussions, assessors praised **Lesson Planning** for its ability to elicit information about a teacher's pedagogical knowledge and pedagogical content knowledge, especially with regard to a teacher's design of instruction. The assessor who administered **Topic Sequencing** felt that the task based on elementary fractions was reasonable, while the one based on ratio and proportions "did not work." Other assessors and scorers felt that many elementary teachers are not prepared to be tested on their knowledge of mathematical concepts and how they interrelate. Assessors felt that there were alternative models of tests such as existing multiple-choice tests that adequately examine a teacher's content knowledge, and that interviews were more appropriate for assessing pedagogy and pedagogical content knowledge.

On the whole, assessors and scorers felt that the method of semi-structured interviewing had some merit, although the SSI-EM itself needed revisions to address pedagogy and pedagogical content knowledge more fully. A couple of the assessors and scorers agreed with many teachers in their belief that simulations and artificial conditions did not fully tap a person's ability to teach and that any interviews should be supplemented by classroom observations. One scorer believed that some of the poor-scoring teachers were probably good teachers in other subjects. The two assessors on sabbatical from classroom teaching felt that supplementary materials should be provided to make the assessment more reflective of teaching by first-year teachers.

Although the SSI-EM demands a high level of content knowledge, it assesses quite well a teacher's ability to represent content, explain concepts, and sequence instruction. All of these are aspects of content pedagogy which depend on a sophisticated level of content knowledge.

Assessment Format

Teachers, assessors, and scorers were asked their perceptions of various aspects of the Semi-Structured Interview and Assessment Center formats. Their comments are summarized below. This section provides information about the clarity of the assessment, the choice of tasks, the use of probes, and the use of interviewers.

Clarity of Assessment

A detailed description of teacher responses is found in the *Administration Report for Spring, 1989.* Teachers generally felt that the materials and instructions for the SSI-EM were clear. Roughly two-thirds of the teachers found the written materials they received prior to the assessment to be helpful. These materials did ***not*** include detailed descriptions of each task. Descriptions of the tasks and scoring criteria prior to the assessment would assist teachers in anticipating and preparing for the assessment. It is unclear, however, whether and if so, how, this additional information would affect the level of anxiety experienced by the teachers, especially those who are not confident of themselves in mathematics.

Nearly all of the 40 teachers found the oral overview at the beginning of the day to be helpful, and the directions for the assessment clear. Suggested changes included: send the orientation materials in advance, give a clearer idea of the all-day process, include driving time in the directions, and revise the letter to make it sound less intimidating. None of the teachers suggested changing the directions for the tasks.

Assessors felt that the instructions for the tasks were generally complete, detailed and clear. Few changes were suggested. One assessor suggested that some of the terms, e.g., "assessment instrument," "math concept," were confusing and should be explained or written in the vernacular. Another assessor felt that changes in tone from, "I now want you to..." to "Please, now..." would improve the atmosphere and put the teachers more at ease.

Some teachers experienced difficulties in performing the tasks. Some of the difficulties were at least partially due to lack of content knowledge, e.g., the expressed need for more "background experience in math" to properly prescribe remedial instruction for the **Instructional Vignettes**, and insufficient understanding of some of the topics in **Topic Sequencing**. Other sources of confusion were the lack of information about the evaluation criteria, the redundancy and difficulty of the questions, and a belief that the **Topic Sequencing** tasks depend on the characteristics of the students being taught. We believe that teachers should be informed of the general scoring criteria prior to the assessment and be guided in some manner during the interview to provide responses in sufficient scope and detail to reduce ambiguity in coding their responses. The latter can be done by adding or revising questions to better focus responses, or by giving each teacher a list of areas to be addressed in the response to broad questions (as was done in the SSI-SM), e.g., when providing an overview of the lesson.

The questions sometimes seemed redundant to teachers when they were asked to explain how they would teach a student to work a problem, followed by a request either to explain how the solution worked mathematically or to identify the underlying math principles. Teachers often did not distinguish between algorithms to solve problems and mathematical principles underlying algorithms. In these cases, the questions asking how to solve problems seemed identical with questions that asked teachers to explain why the given solution would work.

With regard to the belief that the sequence of topics varies according to student characteristics, **Topic Sequencing** did not have a single correct answer. Given the way that most teachers had been trained in mathematics to look for the solution and not for alternative approaches, teachers might have anticipated that there was only one correct answer. Although there was no one correct sequence, topics ***could*** be grouped in sets and ordered. Many of the topics to be ordered were prerequisites for understanding other topics. Therefore, some orderings would be incorrect for any group of learners.

Teachers and assessors gave feedback regarding the assessment before the scoring criteria were developed. Although the scoring criteria do not directly correspond to the questions, there ***is*** a rough match between criteria and questions which ***ought*** to yield information addressing specific criteria. For many of the questions, a slight rewording would have made the question more likely to yield a response that directly addressed a scoring criterion. For example, one scoring criterion for **Topic Sequencing** is the validity of analogies and metaphors used to explain concepts to students. None of the questions, however, directly asked teachers how they would explain the topics to their students. Instead, teachers were asked to tell what each topic meant to them. When a response addressing a scoring criterion was ambiguous and the question had not been asked directly, scorers were directed to not score that category and exclude it from the calculation of summative scores.

Several scorers expressed frustration with rating borderline responses, suggesting a need for improvement in distinguishing suitable from unsuitable answers. Their frustration was reflected in the number of responses which they rated as ambiguous, which ranged from seven in **Lesson Planning** to 66 in **Instructional Vignettes**. Some of the scorers coded answers as ambiguous more often than others, contributing to problems in achieving reliability. Nearly all scorers expressed frustration with scoring criteria which did not closely correspond to questions asked. Since scores were for the most part calculated on the percentage of responses judged adequate, excluding a category means that the knowledge of skills assessed by that category are not reflected in the overall score for a task.

The SSI-EM task questions should be revised to better match scoring criteria before the assessment is administered again. These revisions might range from rewording questions so the focus is more explicit to inventing new questions that could elicit a specific type of response. Another route to reducing ambiguous responses is to provide the scoring criteria to teachers in advance so they better frame their responses. The provision of scoring criteria to the assessors should improve their ability to probe more effectively, and suggested probes can be provided for specific questions to standardize the process.

In addition to providing better alignment between questions and scoring criteria, rewording the existing questions for a more explicit focus could also help standardize responses to make them easier to evaluate. One assessor commented that the vignettes produced a wide variety of interpretations, with different teachers seizing on differing but, in her opinion, equally important aspects of the vignette in their response.

Format Features

In the pilot testing of the SSI-EM, several features of the semi-structured interview format were identified as either helpful or problematic. They include the timing of the exercises, the choice of tasks, the use of probes, and the use of interviewers.

Timing. The assessment was originally scheduled to take a little over six hours (including orientation, an hour for lunch, and feedback from teachers). In this schedule, an hour was allocated for each task. In practice, however, teachers did not take the entire time allotted. **Lesson Planning** generally took an hour, but most teachers completed each of the three other tasks in a half hour or less. Although it was known from the outset that **Lesson Planning** would take more time than the other tasks, the extent of the difference was not known.

When the extent of the disparity in length of time required to complete each task was discovered on the first day, assessors began to experiment with ways to accommodate these differences. In the first week of administration, teachers tended to spend the wait time between tasks either reading newspapers or in conversation with assessors in the hospitality room. By the middle of the second week of administration, assessors began altering the schedule by administering a task to a teacher as soon as that teacher had completed the previous task scheduled and had indicated a willingness to continue. This resulted in teachers completing the assessment at varying times, with the teacher who had **Lesson Planning** scheduled last taking the most time.

By the final round of administration, FWL staff had learned from experience how long each task took, so an additional assessor was added to administer **Lesson Planning**, and a schedule minimizing wait time was devised. For an assessment center format to be efficiently implemented, it is essential that tasks be designed to take roughly equivalent amounts of time. In practice, there will be individual differences in task performance that will complicate strict adherence to any predetermined schedule. However, prior piloting of tasks should reveal any great disparities in the average amount of time required for completion.

About three-fourths of the 40 teachers did not feel they needed more time for any tasks, though about one-fifth did, with one teacher mentioning specifically **Lesson Planning** and another **Instructional Vignettes.** Generally, most did not feel there should be less time for of the any tasks.

Choice of tasks. Although the relative length of the tasks needs improving, the four tasks (**Lesson Planning, Topic Sequencing, Instructional Vignettes,** and **Short Cuts**) do a good job of reflecting key activities which teachers must do to effectively plan and manage instruction at a conceptual level. **Topic Sequencing** and **Short Cuts** give good illustrations of the command of the subject matter. All four, but especially **Lesson Planning** and **Instructional Vignettes,** address the ability to translate concepts into appropriate metaphors or activities. **Lesson Planning** and **Topic Sequencing** address the ability to sequence instruction so that concepts are taught in such a manner that they build on previous student knowledge and lay the foundation for material that occurs at a later point in the curriculum.

Questions about one of the tasks, **Instructional Vignettes**, arose during the administration and scoring of the SSI-EM. Our concerns about vignettes are based not only on our experience with the SSI-EM, but also with observation of the use of vignettes in the Stanford BIOTAP pilot test. Few of the teachers had difficulty responding to the vignettes. However, the concerns of teachers who found the vignettes artificial suggested two potential problems in the use of vignettes. The first is that teachers use a variety of cues in formulating a response to students that are unavailable in vignettes. With the exception of the period at the beginning of the school year, teachers know something about their students. In responding to a student, a teacher takes into consideration the student's past level of performance, personality traits such as perseverance, knowledge of what the student has been taught previously in the year and previously established routines for remediation. This information is unavailable to teachers when vignettes are presented, resulting in a variety of responses related to unarticulated assumptions based on the teacher's own experience. Teachers varied in their comfort with making these assumptions; it is possible that teachers who are especially skilled at tailoring responses to individual students might have the greatest difficulty in responding to hypothetical students.

Some of these difficulties can be overcome by adding additional information to the vignette and by piloting the vignette with teachers from varying contexts. For instance, one assessor found that teachers exhibited varying but equally valid interpretations of a single vignette. She recommended that the questions be reworded to make the focal point clear. If a particular focus is desired, then the critical information can be included in the vignette.

Since vignettes *are* artificial, they are unlikely to capture the ability to respond to individual students in a holistic manner which takes into account a student's mood at the time, personality characteristics, preferred learning style, and content knowledge. They *do* seem to capture the ability to create metaphors, use alternative representations of content, and spontaneously design activities to explain concepts and correct misconceptions.

The second problem with vignettes that became obvious during **Lesson Planning** is that teachers vary in the way that they design instruction. Some situations that are designed to elicit teachers' responses to common student errors are inappropriate when the teacher has carefully constructed the lesson to avoid producing such errors. One example is when a student is shown converting 2/50 to a percentage and arriving at an answer of 2%. Some teachers had spent quite some time initially in the lesson talking about the meaning of a percentage, having everyone draw pie charts representing percentages, and converting fractions with 100 as a denominator to percentages. One teacher went so far as to call the process of conversion of a fraction to a percentage "renaming." If, as in these instances, a teacher has carefully laid a foundation for student understanding that makes it less likely that students make this type of error, then the only situations in which students produce that type of error are (1) if they have completely missed the point of the lesson, or (2) when they are displaying careless thinking. Teachers who make these assumptions will react differently from teachers who assume this is an instance of an incorrect algorithm applied by the student which produces a consistent pattern of errors. If the scoring system assumes a single source of student error or fails to probe for assumptions about the source of error, the scoring may fail to reflect the more complex teaching.

As a whole, the SSI-EM gives a good picture of what Lee Shulman calls "pedagogical content knowledge." However, the methodology of semi-structured interviews is still in need of improvement, particularly in the vignettes (discussed earlier) and the construction of probes.

Use of probes. The purpose of probes was to assist in the interpretation of ambiguous responses. Responses could be ambiguous because of their brevity or because of the use of educational jargon. Without elaboration, it was often difficult to distinguish a terse summary of a complex thought from a vague explanation which lacked depth of analysis. Assessors were carefully trained and monitored in practice administration on how to construct probes which were clarifying, and which did not cue the teacher as to the appropriate response. In practice, however, assessors found the use of probes problematic for several reasons. The first is that some of the questions were unclear or used terminology with which teachers appeared unfamiliar. An example of a type of question that was unclear was one asking the teachers to explain how they might integrate instruction on how to use a particular mathematical algorithm into their teaching when they had in the previous question explained why they found it unsuitable. It was difficult to probe or even to respond to a teacher's questions if the assessor was not clear about the intent of the question. Some teachers had difficulty responding to questions asking about "mathematical concepts" or "mathematical reasoning," even after they had just described accurately how to work the problem.

The second reason probes were problematic is the difficulty mentioned above of avoiding cues to the teacher about the correct answer. In the questions referred to above, which asked about "mathematical concepts," it was particularly difficult for an assessor to adequately explain the intent of the questions without giving cues about the appropriate response. Scorers reported that assessors sometimes gave in to the temptation to lead the teacher to the correct answer, particularly when the teacher seemed to be slowly progressing toward the correct answer.

The third problem with probes is one of standardization. Several assessors indicated that probing was the most difficult aspect of administration, not only because it was difficult to use probes that were not leading, but tailoring probes to the teachers caused some assessors to question the consistency with which they were using probes. For example, because some teachers were clearly more nervous than others, the assessors reported probing more gently or not at all when a teacher's tone and body language indicated that the probes were increasing the teacher's frustration rather than facilitating construction of an answer. This may have put the more nervous teachers at a disadvantage relative to other more confident teachers.

Inconsistency in the use of probes could be ameliorated by revising the questions so they more closely correspond to scoring criteria, and by improving assessor training. Other problems resulting from the use of interviewers may not be solved as easily, however.

Use of interviewers. One can envision an alternate form of this assessment which asks the teachers to perform the same tasks and then respond to questions in a written format. Although teachers were not asked to comment specifically on the use of interviewers, many of the teachers did so either during informal discussions with the interviewers or on the evaluation feedback forms. On the positive side, some teachers enjoyed the interaction with an interviewer, typified by the teacher who appreciated the opportunity to "think about and reflect on my newly gained teaching techniques." These teachers seemed to enjoy the interaction with the interviewers, and did not feel intimidated.

Other teachers, however, in both our pilot test and other studies (e.g., see Wilson, 1988) were uncomfortable being assessed through an interview. For instance, some teachers mentioned the difficulty they had working one-on-one with someone in an artificial situation (i.e, an interview at an assessment center). One teacher commented that "good teachers don't always interview well and vice versa." Another teacher explained:

> *I don't work well under pressure, in a one-on-one situation. The tasks would have been no problem if I worked them out at home or in the classroom.*

This teacher was not the only one who found the assessment extremely stressful, despite the best efforts of the assessors to put the teachers at ease and the relatively informal atmosphere. The most extreme example of stress was one teacher who, after struggling with the three tasks generally perceived to be the most difficult (**Topic Sequencing, Instructional Vignettes**, and **Short Cuts**), had a strong emotional reaction to the interview and required over an hour to regain composure. While this was the only such instance among the 41 teachers, many teachers commented that their anxiety would be heightened considerably if they were participating in the assessment for credentialing purposes.

At least some of the stress experienced by teachers was probably due to anxieties which are intensified by having another person witness your struggles. Several teachers referred specifically to feelings of "inadequacy," "lowered self-confidence," and "incompetence" they experienced when they had difficulty answering some of the questions. Other teachers, sometimes attending the ***same*** assessment administration as the teachers reporting high anxiety, reported feeling nervous initially, but the assessors were able to put them at ease.

Assessors varied in their tone and degree of formality. One scorer described the range as from "very formal, cold and rather tense" to "relaxed, cordial and even playful with the candidates." While the tone can be standardized somewhat through training, the anxiety which some candidates experience throughout the assessment regardless of assessor style is less easy to address.

It is unclear why teachers in the SSI-EM expressed discomfort while none was reported or observed for the SSI-SM, which was similar in format. One possible explanation is the differing target groups for which the SSI-EM and SSI-SM were designed. The SSI-EM was originally designed for experienced teachers, while the SSI-SM was tailored for beginning teachers. Another possibility is that elementary teachers feel greater anxiety because they are not subject matter specialists, compared to the secondary teachers, who had studied the discipline in which they were being assessed. Our impression is that the SSI-EM assessors were also more aggressive in probing ambiguous answers than those in the SSI-SM, providing numerous cues to teachers who were not doing well.

Cost Analysis

Cost projections for semi-structured interviews were discussed in Chapter 4 in connection with our experience pilot testing the SSI-SM. The costs for piloting the SSI-SM were used since that assessment represented a later stage of development than the SSI-EM, resulting in fewer implementation problems.

Technical Quality

This section discusses three aspects of the technical quality of the SSI-EM: development, reliability and validity.

Development

The SSI-EM was based on four tasks that were developed, pilot tested, and field tested by the Stanford Teacher Assessment Project (TAP). Minor modifications were made in the interview protocols; major changes were made in the scoring criteria. The original set of tasks focused on the topic of elementary fractions, particularly the simplification of fractions. Second versions of two of the tasks, **Lesson Planning** and **Topic Sequencing**, were developed to determine the feasibility of using existing tasks as shells to apply to new content.

The original tasks were developed by the Stanford TAP over a one-year period and were evaluated through a series of activities. First, the interview protocols were piloted with "expert" teachers, mathematicians and mathematics educators. Second, each protocol was critiqued by groups of "expert" teachers. Third, the instruments were similarly reviewed by TAP's Expert Panel in mathematics, composed of mathematics educators, mathematicians, teacher educators, teachers, and TAP staff.

Scoring instruments also went through a similar one-year, multi-staged development and revision process. First, TAP staff devised and tested various scoring formats, resulting in an eclectic set of 10 different scoring mechanisms; some were holistic, while others were analytic. After these scoring procedures were applied to a few sample protocols, their effectiveness was examined by a board composed of teachers, teacher educators, researchers, and TAP staff. After revisions, the scoring procedures were sent to a second board of examiners with a similar composition and once again revised. Teachers (who were not project staff) were trained to use this final set of scoring procedures to score the data collected at the TAP Assessment Center. The TAP staff collected feedback and analyzed these external scorings.

The TAP instruments were revised for the SSI-EM in two ways. First, a TAP representative revised the interview protocols and scoring systems based on feedback from those who took, administered, scored, and analyzed the instruments. Second, the interview protocols were revised to make them more appropriate for beginning teachers. The purpose of the TAP instruments was to identify exemplary experienced teachers. To adapt the TAP instruments for beginning teachers, the TAP representative simplified the tasks and revised the scoring criteria to reflect a less sophisticated level of expertise. Due to time and budget constraints, these revisions were not subjected to wide review.

Reliability

Data on inter-rater reliability and problems encountered in scoring suggest the need for lengthening and strengthening the training of assessors and scorers. Assessors need to be familiar with the scoring criteria to effectively conduct the interview. Scorers need more training in rating performances, especially at the cutoff point.

Two versions of tasks that addressed two different topics were piloted. In general, the questions in the interview protocols were similar and scoring categories were parallel. No teachers performed both versions of the same task, so data on the reliability of the SSI-EM across tasks is not available.

Validity

The interview questions and scoring categories for the SSI-EM focus on very specific aspects of a teacher's performance. Although the interview questions have undergone some review during development, both they and the scoring criteria would need to be subjected to a wider review in order to ascertain whether or not they reflect a broad professional consensus about important teaching competencies. This validity study should be completed before extensive developmental work is done on the SSI-EM or any other semi-structured interview. Since the SSI-EM scoring criteria are very specific, singling out particular aspects of a teacher's response, they would be especially vulnerable to professional disagreements about important components of responses and/or the relative importance of these components. Some of the scorers expressed reservations about the comprehensiveness of some of the scoring categories that used a check list format.

Conclusions and Recommendations

This section contains conclusions and recommendations regarding the SSI-EM, organized into the areas of administration, scoring, content, format, and a brief summary.

Administration of Assessment

Like the other semi-structured interview that was pilot tested, the SSI-EM is very labor intensive to administer; administration and scoring require one day per teacher. If on-line scoring is developed and found to be feasible, the time required for administration could be reduced slightly, but our experience with the CCI suggests that forming and documenting judgments is a time-consuming process.

The following factors seem to be key to smooth implementation of the SSI-EM or any other semi-structured interview assessment:

- o availability of appropriate facilities (which are often difficult to locate);
- o development of clear orientation materials for teachers, including descriptions of the tasks and scoring criteria;
- o organization of tasks so all tasks require approximately equal amounts of time and only a minimal number of transitions are needed between tasks;
- o careful attention to the recording quality of the tape recorders used;
- o coordination and management of a large number of materials and pieces of equipment, including interview protocols, tape recorders, and labeled tapes; and
- o recruitment of assessors and scorers who are knowledgeable about mathematics, mathematics pedagogy, and student characteristics.

Audiotaping proved adequate for documentation of the interviews. Checking the recording quality before each interview seemed to minimize the chances of unrecognized equipment failure; some recording problems were detected in the pilot test, and alternative equipment was used. Precautions need to be taken to minimize the chances that assessors forget to turn on the tape recorder, such as oversized reminders printed in the interview protocol. A policy of recording the entire interview, including the introductory portion and those times when candidates are thinking and not speaking, would minimize the chances of recording only a portion of the interview.

Assessor training should include familiarization with both the scoring criteria and the recording equipment. Some controlled experimentation needs to be done with probes to identify the kinds of situations in which probes are needed for scoring purposes, the effects on performance when probes are and are not used, and variance in assessors' use of probes. Such a study would inform the development of guidelines for standardized use of probes.

As with the SSI-SM, the security needs of the semi-structured interview should be studied to determine its robustness with respect to the development of standardized responses.

Scoring

The scoring system of the SSI-EM is not suitable for adoption for a statewide assessment. Scoring criteria that vary by task and by topic raise serious concerns about reliability, validity and fairness across differing versions of the assessment. The scoring system developed for the SSI-SM is a more promising prototype which avoids these problems.

Our experience with the implementation of scoring criteria which were developed after the administration of the assessment underscores the importance of developing tasks and scoring criteria simultaneously and analyzing their alignment prior to pilot testing.

Scorer training should include clear examples of responses in each rating category. Much time in training should be devoted to differentiating borderline responses, especially at the cutoff point between passing and failing.

Assessment Content

Our observations and information collected from assessors, scorers and teachers participating in the pilot test suggest the following conclusions about the content of the SSI-EM:

- Assessment content was in line with the philosophy of the *Mathematics Framework*. Congruence was good, though not complete, with respect to areas of emphasis and characteristics of delivery of instruction.
- Coverage of the California Standards for Beginning Teachers could be improved by refining current tasks and developing new ones. Some standards covering teacher-student interaction or student outcomes could only be indirectly addressed with this assessment.
- For the most part, the tasks were perceived by teachers as being job-related. The major exception was **Topic Sequencing**.
- The content was difficult for the beginning teachers who participated in the pilot test. Less than one-third of the teachers passed more than half of the tasks. Some difficulties indicated a need for improvement in teacher preparation; others were more likely the result of inexperience, either in teaching in general or in teaching the particular topic.
- Teaching diverse students in a variety of contexts was not addressed; teachers of low-achieving students felt disadvantaged with respect to the questions and tasks.
- Since the targeted teaching competencies were generally highly valued, the SSI-EM was judged as a fair way to assess various teacher groups; pilot test data were too limited, however, to draw conclusions about the performance of minority teachers.

If the SSI-EM is chosen for further development, it would benefit from the same type of content review by an expert panel that was recommended for the SSI-SM.

Assessment Format

Based on experience with the SSI-EM as a state-of-the-art prototype, the strengths of the semi-structured interview format appear to be in assessing the ability to plan instruction and the command of the subject at a conceptual level of understanding. It appears weakest in assessing a teacher's ability to implement instruction and manage the classroom.

Our experience in implementing the SSI-EM suggests that the following issues be considered when contemplating adoption of the semi-structured interview format for teacher assessment:

- On numerous occasions, a teacher's lack of content knowledge affected his/her ability to respond to questions and made it difficult to clarify questions without revealing the nature of a correct response. Questions that were subtly different were seen as equivalent to teachers who did not comprehend the subtle differences.

- Vignettes need to be carefully constructed with a specific focus so all relevant information can be included and the range of possible interpretations is narrowed. Teacher assumptions that may affect the evaluation of their responses need to be explored in the interview.

- The use of interviewers to collect data seems to heighten anxiety for many teachers, especially those teachers who are not performing well.

In general, the use of interviewing, as compared to written or dictated responses to printed questions, should be explored further. The extent to which differences in interviewing style affect teacher performance should be a key consideration.

Summary

If semi-structured interviews are selected as a method of assessing new teachers for credentialing purposes, a close review or study of the information yielded by the SSI-EM tasks and questions could inform the development of prototypes. However, the SSI-EM scoring system does not appear to be a promising approach that bears further development.

CHAPTER 6:

ELEMENTARY EDUCATION EXAMINATION

CHAPTER 6:

ELEMENTARY EDUCATION EXAMINATION

The Elementary Education Examination is an innovative multiple-choice test developed by IOX Assessment Associates for use by the State of Connecticut in the licensure of elementary school teachers, for grades K-8. The examination is part of a three-tier assessment system currently under development by Connecticut. Under this proposed system, the first assessment, a basic skills test in reading, writing, and mathematics, is administered during prospective teachers' undergraduate programs. Prospective elementary teachers who pass the basic skills test and successfully complete an undergraduate degree and a teacher education program must then pass the Elementary Education Examination. Candidates who pass the examination are given an initial teaching certificate and can begin teaching. During their first year of teaching, they are further evaluated through observation and/or performance assessments. Those who pass this third round of assessment are then awarded a provisional teaching certificate.

The Elementary Education Examination focuses on three major competencies: mastery of content knowledge, mastery of pedagogical knowledge, and mastery of pedagogical content knowledge. The examination differs from more traditional multiple-choice tests in two respects. First, the majority of questions -- regardless of the competency area being assessed -- are embedded in classroom situations (e.g., "You are planning a lesson on chemical changes. Which of the following....?"). Second, some of the items (referred to as "materials-based items") ask the candidate to analyze reference materials that are commonly used by classroom teachers. These materials include Individual Education Plans (IEPs), student worksheets (some blank, some with student work), lesson plans, report cards, and test reports. A description of a sample "materials-based item" is as follows:

> Given four worksheets of student learning activities, the examinee must identify the worksheet that most closely matches a specified objective.

The administrative format of this assessment is the same used for other large-scale multiple-choice examinations: Candidates come together at a test site and are assessed individually by their written responses to a series of multiple-choice items. At each administration of this assessment, six different forms of the exam were used in order to pilot test a greater number of items. Each form consisted of 77 multiple-choice items, with 17 items appearing on all six forms. In instances of differences in performance across forms, the 17 linking items serve as indicators as to whether the differences are due to the difficulty of the items or to differences in the ability levels of the examinees. Although two hours was the suggested time for the examination, time limits were not established during this pilot phase.

Beginning with information on the administration of the assessment, this chapter continues with a discussion of the content and the format. Following these discussions are analyses of the cost and technical quality of the assessment. The chapter concludes with an overall summary, together with recommendations for further steps in exploring the feasibility and utility of an innovative multiple-choice examination such as the Elementary Education Examination in California teacher assessment.

Administration of Assessment

This section begins with an overview of the administration of the assessment. It is followed by information on the following: logistics (e.g., development of orientation materials, identification of teacher samples, scheduling), security arrangements, assessors, scoring, and teacher and FWL and RMC staff perceptions of the administration.

Overview

The Elementary Education Examination was administered at nine sites to both project and nonproject teachers. Table 6.1 contains information about the pilot testing of the assessment. Over 250 teachers from a total of nine projects, plus over 300 teachers from 11 nonproject districts, were invited to participate in the assessment. Ten administrations were scheduled between May 4, 1989, and June 13, 1989, all but one of which were scheduled for the late afternoon from 4:00 to 6:00 p.m. One administration was scheduled in two shifts on a Saturday morning, and one was canceled by IOX Assessment Associates due to an anticipated poor turnout (only six out of 16 teachers had said they would come) and a lack of staff to administer the examination.

In all, a total of 138 teachers participated in the Elementary Education Examination assessment (or approximately 25% of the teachers who were invited to participate). Based on information taken from the teacher feedback forms completed by 137 of the teachers, 121 (88%) females and 14 (10%) males were assessed (two respondents did not specify gender). In addition, 76% (105) of the teachers described themselves as White (or Anglo or Caucasian), 12% (17) as Hispanic (or Chicano or Latino), and 3% (4) as Black. The ethnic breakdown of the remaining 8% of the teachers is as follows: American Indian (3), Asian (2), Pacific Islander (1), Other (3), and Not Reported (2).

Logistics

Logistical activities for this assessment included the development of orientation materials, identification of teacher samples, scheduling the test administrations, making site/facilities arrangements, arranging for the assessment materials, developing evaluation feedback forms and securing the evaluation feedback, and reimbursing the teacher participants.

The orientation materials developed for this assessment were very important because they were the means by which teachers were invited to participate in this assessment. These materials included a letter which described the pilot testing project and the Elementary Education Examination, and specified the date, time, and location of the assessment administration. The letter also informed teachers that they would receive $25.00 for their participation as well as mileage expenses if they had to travel more than 15 miles to the test site. Attached to this letter was a self-addressed, stamped postcard

TABLE 6.1

ELEMENTARY EDUCATION EXAMINATION: PILOT TEST PARTICIPANTS

(Total Number of Teachers = 138)

Date	Project	Number of Teachers Invited		Number of Teachers Who Took Test
		Project	Nonproject	
May 6	Santa Barbara/ Ventura	30	–	9
May 20	Riverside/ San Bernadino	119	–	30
May 30	Santa Cruz	36	7	34
May 31	Santa Clara	9	98+*	8
June 1	Centralia	12**	65+*	8
June 6	Long Beach	7	86***	12
June 7	El Cajon	25	25	26
June 8	Poway	18	–	6
June 13	Vista****	–	22	5
		======	======	==========
		256	303	138

*Some districts invited their 1st-year teachers but did not release their names to us.

**1 Centralia Project teacher and 11 Irvine Project teachers.

***Includes 45 Lynwood District teachers who were also invited to the Centralia assessment.

****Nonproject district.

that the teachers were asked to return informing us whether or not they were able to participate in the pilot test.

All teachers selected for the sample for this assessment were sent the orientation materials. This sample included (1) all elementary teachers from the selected California New Teacher Projects who were not scheduled to participate in other pilot tests or any of the Project evaluation activities being conducted by SWRL, and (2) beginning teachers from neighboring Nonproject districts.

After consulting with Project Directors about optimal times, the administrations were primarily scheduled for weekday afternoons. Although original plans called for two or more Projects to participate at each test site, FWL staff found it difficult to find sites that were both geographically centralized and required minimal teacher travel time. Hence, the majority of administrations included only a single Project and its neighboring districts. All administrations were conducted in a school auditorium, cafeteria, or classroom.

Upon completion of the assessment, each teacher was asked to complete an evaluation feedback form and to sign a list which served to verify participation in the assessment. All teachers who signed the list were then mailed a check for $25.00 plus mileage costs if they had to travel more than 30 miles to participate.

Security

IOX Assessment Associates assumed full responsibility for all security arrangements of the test materials and test administrations. The test booklets were numbered, and all booklets were logged in when they were returned by the teachers with the completed answer sheets. Different forms of the test were distributed, minimizing the opportunities for one person to copy another's responses.

Each administration of the examination was supervised by an IOX representative; at no administration, however, was anyone designated as a proctor. For any future administrations in California, a proctor should be present to assist in the test administration and ensure security. Detailed instructions on the security of test materials should be available to the proctor, as well as instructions as to placement of test materials during the administration, counts needed at various stages, and actions to be taken if any materials are missing.

For this pilot test, scrap paper was permitted during the administrations and passed out with the test books and answer sheets. Examinees were also allowed to request additional paper if they needed it. This practice poses a potential security problem. How is the supervisor or proctor to know how many sheets of scrap paper are given to each examinee and to ensure that every piece of scrap paper is collected? Such a procedure makes it quite easy for examinees to copy items and note important information on the test's contents, and remove the paper from the room. Even though examinees are told not to mark in test books, they sometimes make marks or leave smudges by resting their fingers on certain parts of the stimulus materials or next to certain answers; this may provide information for future examinees using that test book. If the Elementary Education Examination or a similarly innovative multiple-choice test is elected for use in California, we suggest having nonreusable test books and not allowing scrap paper.

Assessors

An IOX Assessment Associates representative based in Southern California was responsible for administering the Elementary Education Examination assessment. At each site, the test administrator gave a standardized oral overview of the assessment, directed the teachers on how to take the exam, distributed the test materials, and collected the materials.

Scoring

The Elementary Education Examination answer sheets were in a machine-scorable format and were scored by the Hacienda/La Puente School District in Southern California. Because we have no information about the quality control procedures employed in the scoring, we cannot comment on this aspect of the administration. Whatever procedures are employed, they must ensure a very high level of accuracy of scoring if the results are to be used in making decisions about credentialing of new teachers, and they must include procedures for dealing with unclear erasures, multiple marks, light marks, incorrectly keyed items, printing errors in test books, and other problems that affect scoring results.

As with any multiple-choice test, a teacher's score for this examination reflects the number of items for which the teacher marked the correct answer. The results for this test were reported in terms of the mean p-value, or in other words, as the percent of examinees marking an item correctly, averaged across all items in each of six subject areas: Human Development and Instructional Methods, Language Arts, Mathematics, Social Studies, Science, and Other (i.e., multicultural, arts, physical education, health, and special education). (See Table 6.2 for the results shown as mean p-values.)

Teacher, FWL, and RMC Staff Impressions of Administration

Approximately 63% (86 of 137) of the teachers who filled out evaluation feedback forms stated that the written orientation materials they received before the assessment were helpful. Teacher suggestions for improving the orientation materials were as follows: be more specific about test objectives; state that specific knowledge will be asked for in content areas; include a sample page from the test booklet; and reduce the length and wordiness of the orientation letters.

An even higher percentage of teachers, 71% (97 of 137) found the arrangements for this assessment (e.g., scheduling, room arrangements, and travel distance to assessment site) to be reasonable. From the teachers who found the arrangements to be unreasonable or who had suggestions for improvement, there were 21 comments about scheduling, specifically about the end-of-year date and/or the afternoon time. Basically, the teachers did not like the test being scheduled during the end of the school year because of the many school activities happening then (e.g. report cards, end-of-year parties or trips), nor did they like being tested after school because they said they were

TABLE 6.2

ELEMENTARY EDUCATION EXAMINATION
SPRING 1989 PILOT TEST RESULTS

SUBJECT AREA	AVERAGE NO. OF ITEMS PER FORM	AVERAGE PERCENTAGE OF ITEMS CORRECT (N=138)
Human Development and Instructional Methods	12	71%
Language Arts	23	73%
Mathematics	18	68%
Social Studies	9	68%
Science	8	71%
Other*	7	78%

*Multicultural Education, Arts, Physical Education, Health, and Special Education too tired and it was "too difficult to concentrate." Some of the teachers who participated in the Saturday morning assessment also commented about the scheduling, saying the Saturday morning time was "inconvenient" because it cut into their classroom planning time.

Other complaints or suggestions for improvement made by teachers were as follows:

Improve facilities -- 10
Increase notification time -- 9
Reduce travel distance -- 7

Of the complaints about facilities, most pertained to the room being too noisy or too warm. Increased notification time was desired by those teachers who did not receive their orientation materials until a day or two before the assessment date. The teachers who wanted reduced travel distance were, for the most part, those teachers who had to travel up to 75 miles to participate in the San Bernardino/Riverside assessment.

Assessing the teachers after a full day of teaching during a very busy and hectic time of year was not an optimal choice in timing, and could be the reason for such low participation levels. Other than low levels of participation, no serious administration problems were experienced.

Assessment Content

As mentioned earlier, the Elementary Education Examination covers three major competencies: (1) content knowledge, (2) pedagogical knowledge, and (3) pedagogical content knowledge. The content areas include nine subjects that are taught in the elementary curriculum: reading/language arts, mathematics, social studies, science, the arts, physical education, health, special education, and multicultural/bilingual education. Pedagogical knowledge includes topics such as human development, classroom management, and student motivation. Pedagogical content knowledge refers to the application of pedagogical principles to specific subject areas, such as those listed above.

The developers of the Elementary Education Examination focused their efforts on identifying the knowledge and skills that are necessary for competence or satisfactory performance, as an elementary teacher. The majority of items on each of the six test forms were designed to assess content knowledge or pedagogical content knowledge. As shown on Table 6.2, however, the average number of items representing each subject area differs. The subject area of reading/language arts is represented by the largest number of items (23-24 items on each form), while math is second (18 items on each form). Within a subject area, the proportion of items which assess content knowledge versus pedagogical content knowledge also differs. This difference reflects the fact that (1) groups of Connecticut teachers and subject-matter specialists established guidelines for item development separately for each subject area, and (2) in some of the subject matter fields (e.g., social studies and science) there is little consensus regarding the best ways to organize and present content to students. Thus, in reading and math, two areas where there is relatively high consensus as to the best ways of applying content to pedagogy, there are proportionately more pedagogical content items than in the areas of social studies and science, two areas in which there is less consensus about content pedagogy. Finally, many of the items that assess content pedagogy are the "materials-based items" which ask the teacher to analyze reference materials that are commonly used on the job (e.g., worksheets, test reports).

The content of the Elementary Education Examination is discussed along the following dimensions:

- Congruence with curriculum guide or framework emphasis;
- Extent of coverage of the California Standards for Beginning Teachers;
- Job-relatedness;
- Appropriateness for beginning teachers;
- Appropriateness across different contexts (e.g., grade levels, subject areas);
- Fairness across groups (e.g., ethnic groups, gender) of teachers;
- Comparison with other similar instruments; and
- Appropriateness of the instrument as a method of assessment.

Except for the first two dimensions which refer to curriculum congruence and standards coverage, the discussions of the remaining dimensions are based on the perspective of the participating teachers as reflected in feedback forms and test results, and observations and analyses by FWL and RMC staff. In addition, because the actual test items are the property of the state of Connecticut, we can only discuss the content in a general way, without referring to or describing specific items.

Congruence with California Curriculum Guides and Frameworks

Having been commissioned by Connecticut, the Elementary Education Examination was not designed to be congruent with California's Model Curriculum Guides and Frameworks. However, because this exam is being evaluated in relation to California's credentialing process, FWL staff looked at the assessment to see in which areas there is congruence with the guides and frameworks and in which areas there is not. In particular, two of the exam's six forms* (Forms #4 and #5) were arbitrarily selected and checked against the *English-Language Arts Guide*, the *Mathematics Framework*, the *Science Guide*, and the *History-Social Science Framework*. In addition, the objectives for the examination listed by IOX Assessment Associates were checked for congruence.

---

*Although all six forms were pilot tested in 1989, one of the forms was developed and pilot tested earlier, in 1988.

As will be evident in the following discussions, the guides and frameworks vary markedly across subject areas in terms of curricular aspects discussed, such as philosophy of instruction, curriculum content at specified grade levels, and desired characteristics of instruction.

The *English-Language Arts Guide* has 22 guidelines categorized into five major groupings. The first grouping emphasizes the reading and the study of significant literary works. Although one of the exam's 11 reading/language arts objectives mentions "the use...of children's literature selections," FWL staff's analysis of the 20-24 reading/language arts items on each of the two test forms did not reveal any items that dealt specifically with a literature-based reading program. Since a major thrust of California's reading/language arts curriculum is a literature-based program, the exam would need revisions to address this area. The second grouping emphasizes classroom instruction based on students' experiences. Our analysis again revealed a lack of items and objectives corresponding to this grouping. The third grouping refers to an interrelated program of listening, speaking, reading, and writing. The majority of items on each analyzed form and the majority of the exam's objectives fall into this category. In particular, many of the items focused on reading comprehension and decoding strategies, as well as on the writing process. The fourth grouping emphasizes a program that is integrated across the curriculum. This emphasis corresponds with one of the exam's objectives, and there are one and three corresponding items on each of the forms analyzed. Finally, the fifth grouping focuses on assessment methods. Three of the exam's 11 objectives correspond to this focus, as do approximately three items on each form analyzed.

The *Mathematics Framework* discusses curricular content and characteristics of instruction. Curricular content is organized into five major emphases: problem solving, calculator technology, computational skills, estimation and mental arithmetic, and computers in mathematics education. The nine mathematics objectives for the exam and the 18 math items on each form analyzed address three of these areas: problem solving, computation, and estimation. None of the objectives or items refer to calculator technology or computers in mathematics education. The bulk of the items refer to computational skills, with a few items addressing problem solving, and only one item on each form analyzed referring to estimation. Ten characteristics of instruction are described by the framework. In the two forms analyzed, six of the characteristics are addressed by test items: Teaching for Understanding, Reinforcement of Concepts and Skills, Problem Solving, Use of Concrete Materials, Corrective Instruction/Remediation, and Mathematical Language. There are no items, however, that address the characteristics of instruction described as Situational Lessons, Flexibility of Instruction, Cooperative Learning Groups, and Questioning and Responding. Should the exam be used in California teacher assessment, consideration should be given to insuring that these characteristics of instruction are reflected in the test objectives and in actual test items on each form.

The *Science Guide* describes science programs for grades K-3, 4-6, and 7-8. Each program is divided into three areas: biological science, earth science, and physical science. Although the eight science items on each of the two forms analyzed cover all three areas, the number of items per subject area is not the same. Form #5, for example, has only one earth science item, but four physical science items. There is also an imbalance in the number of items distributed among grade levels. Almost all of the science items pertain to content knowledge specified in the guide for grades 4-6 or 7-8. On the two forms analyzed, the greatest number of items pertaining to a K-3 program is one. (The other form has one item which may pertain to a K-3 or 4-6 program.) The exam could be improved to ensure greater balance in representation of content areas and grade levels.

The *History-Social Studies Framework* first specifies curriculum goals and strands, and then describes a sequential curriculum for grades K-12. The goals and strands are organized into three broad categories: (1) knowledge and cultural understanding (e.g., historical literacy, geographic literacy, cultural literacy), (2) democratic understanding and civic values (e.g., constitutional heritage, civic rights and responsibilities), and (3) skills attainment and social participation (e.g., basic study skills, critical thinking skills, and participation skills). Each category is addressed by at least one of the eight social studies objectives and by at least one of the eight or nine social studies items on each form analyzed. There is, however, again an imbalance in the number of items distributed among grade levels. On each of the two forms analyzed, only one of the social studies items pertains to grades K-3. For the California context, the exam should strive for greater balance in representation of grade levels, as well as include items that assess knowledge of California history.

In summation, FWL staff would describe the congruence of the Elementary Education Examination with the *English-Language Arts Guide* and the *Mathematics Framework* as fair. Items could be added to the exam to ensure that all major emphases of both the *English Guide* and the *Mathematics Framework* are covered. We would describe the congruence of the exam with the *Science Guide* and *History-Social Studies Framework* as high, but suggest that, in the areas of both science and social studies, there could be a better balance in the number of items pertaining to different grade levels, and, in the area of science, there could be a better balance of items pertaining to the different content areas.

Extent of Coverage of California Standards for Beginning Teachers

Although this assessment was developed for preservice or beginning teachers, it was not developed with the California Standards for Beginning Teachers in mind. Moreover, the California Standards (#22 through 32) define the levels of pedagogical competence and performance that teacher candidates are expected to attain as a condition for earning a credential, while the Elementary Education Examination does not assess performance at all. However, a goal of the Elementary Education Examination is to assess the pedagogical knowledge which the standards represent, FWL staff analyzed the exam's objectives and items designed to assess pedagogical knowledge (described by IOX as knowledge of human development and instructional methods) to see how well they are congruent with the California standards. For Standard 30, which assesses the capacity to teach cross-culturally, we also looked at the objectives and items designed to assess multicultural/bilingual education knowledge. Listed below are brief descriptions of Standards 22 through 32 (the standards appear in italics), accompanied by descriptions of the focus of the test items (based on the exam's objectives) that correspond to the standards.

Standard 22: Student Rapport and Classroom Environment. *Each candidate establishes and sustains a level of student rapport and a classroom environment that promotes learning and equity, and that fosters mutual respect among the persons in a class.* This standard is not addressed directly, but some test items assess a teacher's knowledge of methods that enhance students' self-concepts, increase students' motivation, and promote positive attitudes towards learning -- all factors which contribute to rapport with students and help establish the classroom environment.

Standard 23: Curricular and Instructional Planning Skills. *Each candidate prepares at least one unit plan and several lesson plans that include goals, objectives, strategies, activities, materials and assessment plans that are well defined and coordinated with each other.* Some test items require teachers to demonstrate knowledge of appropriate instructional objectives and lesson plans; others require teachers to demonstrate knowledge of appropriate sequences for presenting concepts in an instructional unit and analysis of complex concepts into their constituent parts.

Standard 24: Diverse and Appropriate Teaching. *Each candidate prepares and uses instructional strategies, activities and materials that are appropriate for students with diverse needs, interests and learning styles.* Some test items focus on the selection of learning activities, materials, and explanations based on students' characteristics and the skills and concepts to be learned.

Standard 25: Student Motivation, Involvement and Conduct. *Each candidate motivates and sustains student interest, involvement and appropriate conduct equitably during a variety of class activities.* Some test items assess a teacher's knowledge of methods of increasing students' motivation; others assess a teacher's knowledge of behavior management and classroom management strategies that promote prosocial behavior.

Standard 26: Presentation Skills. *Each candidate communicates effectively by presenting ideas and instructions clearly and meaningfully to students.* None of the test items assess a teacher's ability to communicate effectively with students.

Standard 27: Student Diagnosis, Achievement and Evaluation. *Each candidate identifies students' prior attainments, achieves significant instructional objectives, and evaluates the achievements of the students in a class.* Some test items address the provision of appropriate diagnosis of student difficulties; others are concerned with the appropriate interpretation of data from diagnostic tests and cumulative folders.

Standard 28: Cognitive Outcomes of Teaching. *Each candidate improves the ability of students in a class to evaluate information, think analytically, and reach sound conclusions.* Some items assess a teacher's knowledge of effective questioning strategies -- strategies that improve a student's ability to evaluate information and/or think analytically.

Standard 29: Affective Outcomes of Teaching. *Each candidate fosters positive student attitudes toward the subjects learned, the students themselves, and their capacity to become independent learners.* Some test items require a teacher to demonstrate knowledge of methods of enhancing students' self-concepts and promoting positive attitudes toward learning; other items focus on strategies to encourage students to assume increasing responsibility for themselves.

Standard 30: Capacity to Teach Cross-culturally. *Each candidate demonstrates compatibility with, and ability to teach, students who are different from the candidate. The differences between students and the candidate should include ethnic, cultural,* *gender, linguistic and socioeconomic differences.* Some test items assess a teacher's knowledge of the effects of acquiring English as a second language on students' cognitive and social-emotional development; others assess a teacher's knowledge of cultural differences related to student learning styles and teacher/student interaction.

Standard 31: Readiness for Diverse Responsibilities. *Each candidate teaches students of diverse ages and abilities, and assumes the responsibilities of full-time teachers.* Almost all items on this test are grade-level specific (grades K-8) as reflected by subject matter or student ability (e.g., fourth-grade math or a fifth-grade student reading at the second-grade level). None of the test items, however, addresses a teacher's ability to effectively fulfill a broad range of teaching responsibilities (e.g., preparing for class, meeting school deadlines), which is also a part of this standard.

Standard 32: Professional Obligations. *Each candidate adheres to high standards of professional conduct, cooperates effectively with other adults in the school community, and develops professionally through self-assessment and collegial interactions with other members of the profession.* None of the test items assess a teacher's professional conduct, collegial relations, or professional self-development.

Even though this assessment addresses all but two of the ten California standards, we believe that this assessment is not a particularly effective way to assess a teacher's skills in these areas. Standard 22 (Student Rapport and Classroom Environment) and Standard 25 (Student Motivation, Involvement and Contact), for example, are probably best assessed with a classroom observation instrument. However, the question of whether a multiple-choice test is sufficient to assess basic proficiency in any of the pedagogical areas described by the standards can only be answered by studies of relationships between multiple-choice test scores and scores yielded by more direct measures of skills.

In summation, as shown by Table 6.3, the Elementary Education Examination covers most of the California Standards for Beginning Teachers but does not do so in any depth. Should the exam be revised, we suggest that more items could be included that address in greater depth each standard now covered, but that ***first priority*** should be the addition of items that improve the exam's capability of assessing a teacher's capacity to teach cross-culturally (Standard 30) and of evaluating a teacher's ability to teach students with diverse needs, interests, and learning styles (Standard 24). (The addition of items would, we realize, extend the length of the test, a tradeoff that may or may not be acceptable.)

Job-relatedness

Teachers who participated in this assessment were asked if they felt that the subject areas and concepts chosen for this assessment are relevant to their job of teaching. Approximately 60% (82 of 137) responded affirmatively. Many of these teachers, however, qualified their "yes" answer, responding that some of the questions were more relevant than others. Certain math or science questions, for example, were cited by some teachers as being irrelevant to their jobs.

Twenty teachers specifically commented that the subject areas and concepts chosen for this assessment are not relevant to their job of teaching because they teach at a grade level for which many of these subject areas and concepts are not applicable.

TABLE 6.3

EXTENT OF COVERAGE BY THE ELEMENTARY EDUCATION EXAMINATION OF CALIFORNIA STANDARDS FOR BEGINNING TEACHERS

Standard	Content Focus of Test Items and Objectives Corresponding to Standard	Extent of Coverage
22: Student Rapport and Classroom Environment	-Student's self-concepts, motivation, positive attitudes toward learning	Partial
23: Curricular and Instructional Planning Skills	-Instructional objectives and lesson plans -Sequencing/analysis of instructional concepts	Partial
24: Diverse and Appropriate Teaching	-Learning activities, materials, and explanations based on students' characteristics and skills to be learned	Partial
25: Student Motivation, Involvement, and Conduct	-Student's self-concepts, motivation, positive attitudes toward learning -Behavior/classroom management strategies	Partial
26: Presentation Skills	-Not covered	None
27: Student Diagnosis, Achievement, and Evaluation	-Diagnosis of student difficulties -Interpreting test/cum folder data	Partial
28: Cognitive Outcomes of Teaching	-Questioning strategies	Partial
29: Affective Outcomes of Teaching	-Student's self-concepts, motivation, positive attitudes toward learning -Behavior/classroom management strategies	Partial
30: Capacity to Teach Crossculturally	-ESL instruction -Learning styles to teach/student interaction	Partial
31: Readiness for Diverse Responsibilities	-Not covered	None
32: Professional Obligations	-Not covered	None

Primary-grade teachers especially felt that many of the test items are not relevant to their grade level. One teacher wrote:

> *I am a primary-grade teacher. I have no desire to teach many of the concepts that were on the test. I am certain this has no reflection on me as a first-grade teacher.*

Several teachers suggested that separate tests be developed for primary- and upper-grade elementary teachers. (If separate tests were developed, however, it seems that the current elementary credential covering grades K-8 would no longer be appropriate.)

Because a special feature of this examination is the inclusion of materials-based items which require the teacher to analyze materials commonly used in the classroom, such as lesson plans, IEPs, etc., we asked the teachers how well these items reflect the tasks they perform as teachers. Sixty-six percent of the teachers (91 of 137) responded positively, with answers generally ranging from "ok" to "great." One teacher responded:

> *Materials-based items were quite reflective of my teaching tasks.*

Thirteen teachers stated that the items did not reflect well on the tasks they perform as teachers. Remarked one teacher:

> *They are irrelevant in terms of* ***really knowing*** *the student and their intentions and the materials available.*

Other teachers commented that the items reflected the "ideal" and not the "real" classroom. One teacher, for example, described some of the materials referred to in the items as being inappropriate for her "migrant majority/Limited English Proficient (LEP) classroom."

Appropriateness for Beginning Teachers

When asked if they felt they had sufficient opportunity to acquire the knowledge and ability needed to respond in a reasonable manner to the assessment questions, 83% of the teachers (114 of 137) responded affirmatively and 22 (16%) teachers said "no." (One teacher did not answer.) The majority of teachers also did not find any parts of the assessment to be too easy or too difficult. Of those teachers who did find some parts of the assessment too difficult, the areas of science and math were identified by 22 (16%) and 11 (8%) teachers respectively.

An analysis of the Elementary Education Examination test results (Table 6.2) shows that the percent of items that the teachers marked correctly is roughly equivalent across content areas. If the "other" category (which represents several content areas) is excluded, the average percent of items answered correctly ranges from a low of 68% in mathematics and social studies to a high of 74% in language arts. Thus, even though most teachers felt adequately prepared to "respond in a reasonable manner to the assessment questions," the teachers correctly answered, on average, only two-thirds to three-fourths of the items in each area. Based on these results, there is a distinct possibility that the teachers were insufficiently prepared, either through education or experience, for this assessment.

Appropriateness across Contexts

Fifty percent of the teachers (69 out of 137) assessed believe this assessment is appropriate for teachers in different contexts (i.e, across grade levels, subject areas, and various student groups); 39% (53 teachers) think it is not, and 11% (15 teachers) did not respond at all or gave ambiguous answers. A closer look at appropriateness across grade levels and subject areas and for teachers of diverse student groups is taken below.

Grade level and subject area. Of the teachers who responded negatively to the question of appropriateness across contexts, the majority found the assessment to be inappropriate for teachers across grade levels. There were again many suggestions for developing separate tests for primary-grade teachers and upper-elementary teachers. Although it could well be argued that all elementary school teachers should have mastery of content knowledge and pedagogical knowledge associated with grades K-8 (i.e., the grades corresponding to the elementary credential), it is interesting that many of the teachers do not share this point of view.

Our analysis of the test items found there is a disproportionate number of items representing the subject areas and concepts associated with the upper grades (i.e., grades 4-8), especially in the areas of math, science, and social studies. Of the eight science items on Form #5, for example, there is only one item related to the primary grades, but at least three items specifically geared to grades 7-8 and three for grades 4-6 (it is unclear whether the content of one other item is appropriate for grades 7-8 or grades 4-6). If a single test is used for both primary- and upper-grade elementary teachers, there should be a better balance of items representing the different grade levels.

Diverse students. Seven of the 53 teachers with negative responses deemed the assessment to be inappropriate for teachers of bilingual or LEP students, and three teachers found it inappropriate for teachers of high- or low-ability students. Observed one teacher:

> *In many parts of California, the classroom population reflects a greater proportion of bilingual students as well as lower-achieving students than represented by your test.*

In our reading of the exam's objectives, we found two (out of 61) that pertain to multicultural/bilingual education. One objective is to assess a teacher's knowledge of the effects of acquiring English as a second language on students' cognitive and social-emotional development. The other assesses a teacher's knowledge of learning styles and teacher/student interactions. Although we believe that both of these objectives are good ones (especially in the California context), our analysis of the test items revealed that in some of the forms there are no items that address teaching bilingual/LEP students or students who are characterized as low- or high-ability, and at most there are two items corresponding to these objectives. The test could be improved by adding more items that address the teaching of bilingual/LEP students and low- or high-ability students; however, the identification of low- or high-ability students is often dependent upon the context (i.e., a high-ability student in one classroom may be considered a low-ability student in another classroom). For this reason, items that are developed should carefully define the focal student population. The addition of these types of items would also strengthen the capacity of the Elementary Education Examination to assess teaching competencies encompassed by Standard 24 for teaching diverse students.

Fairness across Groups of Teachers

Seventy-five percent of the teachers (103 of 137) felt this assessment to be fair to new teachers of both genders, different ethnic groups, different language groups, and other groups of new teachers. Fifteen percent (21 teachers) disagreed, and 9% (13 teachers) did not answer or gave an ambiguous answer. Although most of the positive responses consisted of a simple "yes" answer, a few of these answers were qualified as follows:

> *Yes, unless they are only planning to teach very restricted and/or specialized subjects and/or students.*
>
> *Yes, if it's translated for different language groups.*

Most teachers who responded negatively did not explain their answer; seven of the teachers, however, felt the assessment would not be fair to any teacher who was not fully English-proficient.

Appropriateness as Method of Assessment

When asked if they thought that this type of assessment is an appropriate way of assessing teacher competency in the areas of content knowledge, pedagogical knowledge, and pedagogical-content knowledge, 52% of the teachers (72 of 137) responded positively, 40% (55 teachers) responded negatively, and the remaining 8% (10 teachers) either responded ambiguously or did not respond at all. The teachers who responded positively usually did so with a simple "yes." A few teachers, however, elaborated further:

> *Yes, the assessment enabled me to think back to "school days." Often we forget to think about theory, etc.*
>
> *Yes, the assessment forces you to* ***think*** *thoroughly and effectively.*

The teachers who responded negatively were almost always expansive -- often passionately so -- in their response. For example, one teacher wrote:

> *No! Competency of a teacher cannot be made by testing for knowledge alone. A person who scores high on this test may not be as competent as someone who scores lower. It's how the knowledge is used daily in the classroom that counts!*

This sentiment was echoed by many teachers. Some of the teachers indicated that some people don't test well but perform well in the classroom, and others test well but are poor teachers. Some teachers stated that this assessment does a better job assessing a teacher's reading ability than teaching ability. Other teachers expressed the opinion that this assessment is not sufficient to assess a teacher's competency, but should be accompanied by, or replaced with, interviews and/or classroom observations. Even many of the teachers who said "yes," qualified their answer with the proviso that this type of assessment should not be the sole measure of a teacher's competency.

Along the same line, teachers -- particularly primary-grade teachers -- often commented that it was not appropriate to assess a teacher on content knowledge that was not used in their grade level. Numerous teachers suggested that, in order to be fair, separate tests should be created for lower-grade (K-3) and upper-grade (4-6) elementary teachers. For example, one first-grade teacher commented:

> *I think some of the questions can't be completely appreciated by* ***new*** *teachers unless they've taught in a particular grade level. Maybe a test could be made that's more primary-oriented for someone like myself.*

Even some of the upper-grade elementary teachers felt that some of the questions (especially some of the science questions) were more appropriate for middle-school or secondary-school teachers and suggested that a division of tests should be developed accordingly. (As we indicated earlier, our analysis of the test items revealed that a disproportionate number of the eight science items on the two forms analyzed were geared for grades 7-8, or the middle school level.)

For some of the content areas, such as social studies and science, our analysis indicates that the focus of the items is almost exclusively on content knowledge rather than on ***how to teach*** the content (i.e., content pedagogy). The reasons for this can probably be found in our earlier discussion of the constraints experienced by the test developers when designing test items; for example, the lack of general consensus about the application of pedagogy to the content areas of social studies and science made it more difficult to create pedagogical content items in these areas. In the area of reading/language arts, however, there is more agreement in the field about appropriate teaching methods and sequences, and thus there are more reading/language arts pedagogical content items. A large proportion of these items, however, require teachers to correctly match learning activities to a named teaching method. Thus, in the areas of social studies, science, and reading/language arts, the teachers' criticisms of the test, i.e., that it does not necessarily reflect one's ability to teach, were sound. In other areas, such as mathematics, many of the items consisted of activities such as identifying the correct instructional sequence of worksheets to teach a specific concept, identifying concepts whose mastery was necessary to teach a specific new concept, or identifying a pattern of student errors. It would be more difficult to make the case that performance on these types of items is unrelated to teaching competence.

Comparison with Other Multiple-Choice Tests

Because the multiple-choice test is a common method of assessment, the teachers who took this exam were asked to compare it with other multiple-choice exams that they have taken as teachers. Fifty-three teachers gave responses that suggested this assessment is better than other multiple-choice exams they have taken. Many of the teachers specifically compared the test to the NTE and CBEST exams. For the most part, the reasons for the favorable judgments were that this examination assesses more than just content knowledge, and that the test is "more relevant to teaching" and more "teacher-oriented."

Fourteen teachers, however, found the examination to be very similar to other examinations (e.g., "reminded me of the NTE"), and four teachers specifically judged it to be worse. Twenty-four teachers gave no response, and the remaining teachers gave responses that were ambiguous or did not address the question.

Assessment Format

Clearly, the multiple-choice examination is one of the easiest assessments to administer. Large numbers of teachers can be assessed in a relatively short period of time, facilities for administration are generally available, and the number of staff required for each assessment is minimal. The written multiple-choice format measures knowledge of widely accepted principles and basic information, but cannot measure depth of understanding or analytical abilities as well as other assessment approaches. The format also allows for easy scoring. The format issues considered by the teachers who participated in this assessment, however, were those issues specific to this assessment: the clarity of the oral overview and directions presented at the start of the assessment, the clarity of the items, the timing of the test, and teacher preferences regarding feedback.

Clarity of Oral Overview and Directions

Approximately 74% (102 of 137) of the teachers found the oral overview given before the assessment helpful, while 17% (23 teachers) disagreed, and 9% (12 teachers) gave no answer. Six teachers (not all at the same site) said they were not aware that an oral overview had been presented prior to the test. Nearly all the teachers (130) found the directions for this assessment clear; only one teacher disagreed, and six did not state their opinion on this matter.

Should the use of this assessment in California be explored further, we suggest that the directions could be improved by adding information about handicapped examinees, late arrivals, irregularities, or other special circumstances. This additional information should be a part of the field testing as well as of an operational program so that the field test administrations are more similar to the conditions of an operational program and thus yield more realistic data.

Clarity of Items

When asked if they had trouble understanding any of the questions, 64% (88 of 137 teachers) responded "no," and 35% (48 teachers) responded "yes." (One teacher did not respond.) Teachers apparently had the most difficulty with the length and/or wordiness of many of the questions, and with the terminology used in some of the questions. Several teachers also cited difficulty with the many questions that asked for the MOST or LEAST appropriate answer. Other sources of difficulty were the actual content of the questions (e.g., science or math) and the use of reference materials (i.e., the materials-based items). The latter were most often found confusing because of their length (four successive pages of worksheets to review in order to answer one question) or format (the necessity of flipping back and forth between pages to answer the questions). Some teachers also found that for some questions more than one answer seemed correct.

Our analysis of the format of the questions revealed that many of the items are long and wordy, and some, especially the Language Arts items, refer to very specific terminology (e.g., the Cloze method of teaching reading). In the former case, we did not find that the information presented was necessarily extraneous, and in the latter case, we noted that if a teacher did not recognize the terminology, she/he would most likely be unable to answer the question.

We also agree with the teachers who cited difficulty with the MOST/LEAST questions. As FWL staff took the test, we noticed that some of our answers were incorrect because we had neglected to read that the question asked for the MOST or LEAST appropriate answer. Sometimes, if we had answered one or two questions that asked for the MOST appropriate answer, we tended to assume -- incorrectly -- that the next question was asking for the same response. The MOST/LEAST format increased the probability of marking a response that did not accurately reflect a teacher's knowledge. If only a few items were of this format, the difficulty might not be important, but because at least 50% of the items on the two forms analyzed are of this format, the low results may not necessarily reflect a lack of knowledge but possibly an incorrect reading of some or many of the items. This problem could be remedied by limiting the format to either **MOST** or **LEAST** items, but not both.

We also understand why some teachers found the materials-based items to be a source of difficulty. In this case, the difficulty is not in choosing the correct response to the items, but rather the time it takes to read or look at up to four pages of reference materials in order to answer one question. The reference materials comprise at least 33% of the test pages, but only about 15% of the test. For example, on Form #5, 27 of the 80 pages are reference materials corresponding to 12 of the 77 items. Because of the additional reading required, it is probably safe to say that the materials-based items require more time to answer than do the other items. If the test is revised, more items should be added for the longer sets of stimulus materials.

Timing of Tests

Although the pilot test administrations were untimed, teachers were told to expect the examination to take about two hours. Ninety-six percent, or all but six, of the 137 teachers felt that two hours was sufficient time to take the test. Teachers generally finished the test in one and a half to two hours, with a few teachers taking just over an hour.

If a written test of pedagogical knowledge is used in California, we suggest that a time limit be set that permits essentially all examinees to reach the last item on the test. Untimed tests can lead to problems of staffing test centers, locating test facilities, having test materials returned on time if same-day shipments are needed, and increasing anxiety of examinees who do not know how to pace themselves or how much time to take.

Feedback

Although feedback was not given as part of this assessment, teachers were asked what type(s) of feedback would have been helpful. Some of the different answers and the number and percent of teachers who gave them are as follows:

overall score/scores in different areas -- 27 (20%)
weaknesses/areas needing improvement -- 27 (20%)
strengths and weaknesses -- 21 (15%)
the right answers -- 14 (10%)
no feedback -- 10 (7%)

Teachers were also asked to specify by whom the feedback should be given, when, and in what format. Although almost a third of the teachers (43 of 137) did not respond to this question, the answers most often given in response to who should give the feedback were, in order of frequency, the testing company (31 teachers or 22%); someone in the district such as a mentor teacher, a teacher on leave, a principal, or a district support person (22 teachers or 16%); and someone from the university (6 teachers or 4%). Sixteen percent (22 teachers) stated that the feedback should be given as soon as possible, and 16% thought it should be provided in a some sort of written form (e.g., computer printout or narrative).

If a pedagogical knowledge test is used in California teacher assessment, we suggest that examinees be provided with test results as scaled scores that equate the various forms of the test, and with some information on how well they did, possibly in the form of pass, barely fail, and fail rather than percentile rank. With a test covering so many aspects of teaching, the provision of results by types of items is questionable and could lead to misinterpretation of the test results.

Cost Analysis

Rather than use the costs associated with the development and pilot testing of the Elementary Education Exam, we estimate that the cost to administer and score this type of system would be similar to those of other multiple-choice fixed-response assessments such as the CBEST and NTE exams. The costs of these exams currently are in the $32-40 range per teacher. Although the development costs would be somewhat greater for the types of items that are included in the Elementary Education Exam, the administration and scoring costs would be similar to these.

Technical Quality

Item difficulty and correlational data were provided on the Elementary Education Examination. Summaries of these are provided in Appendix C under Construct Validity. The mean comparisons that are provided illustrate the level of difficulty of the different areas in the Elementary Education Examination for different groups. These data demonstrate that questions were of moderate difficulty. Mean p-values (the percent of teachers getting items correct) were in the high 60's and low 70's across subject areas.

Correlational data did not support that the different subtests form clear and separate scales. For example, the math scores on the Elementary Education Examination correlates no better with math scores on the SAT than it does with Language Arts or Social Studies scores on the Elementary Education Examination. These data should be interpreted with caution since the items were in the initial pilot test stages. Other analyses that might be of interest are to separate the "traditional" and "innovative" items to determine whether there is differential performance and information that is attributable to the types of times within each area. Data to perform these analyses were not available to FWL for this report.

The Elementary Education Examination items and specifications should also undergo a content review by California educators. They should examine each item for clarity, accuracy, sensitivity, job relevance, and relationship to the California curriculum frameworks and credentialing requirements. They should look at the test specifications for completeness, clarity, importance for credentialing, and job relevance. All items should also be reviewed by professional editors and sensitivity reviewers, if they have not already undergone such reviews. These reviews should be done prior to the field testing of the Elementary Education Examination if California decides to explore further the possible use of this assessment instrument for credentialing new teachers in California.

Finally, in reviewing the possible use of a pedagogical knowledge test in California, the State must determine what additional information would be provided by this assessment approach above and beyond what is already provided by other tests that are used to make credentialing decisions (e.g., NTE General Knowledge Test), and how much the use of this particular assessment would result in improved credentialing decisions.

Conclusions and Recommendations

This section contains conclusions and recommendations regarding the Elementary Education Examination, organized into the areas of administration, content, format, and a brief summary.

Administration of Assessment

Like other large-scale multiple-choice examinations, the Elementary Education Examination is administered simultaneously to a large number of people. Benefiting from many years' experience in conducting such examinations, the administration of the

Elementary Education Examination poses few logistical problems. The most crucial logistical activity is the selection of the assessment sites. Although the Elementary Education Examination is much less expensive to administer than the other assessments piloted, the economy of scale achieved depends on the number of teachers participating at a single site. Therefore, the higher degree of centralization afforded by this assessment may place larger burdens on teachers from rural areas who will have to travel several hours to a selected site.

The following improvements in test administration are suggested to bring the administration of the Elementary Education Examination more in line with generally accepted practice:

- the use of one or more proctors in addition to the test administrator to monitor test-taking;
- elimination of the use of scrap paper, which poses a security risk; and
- the use of nonreusable test booklets to reduce the incidence of additional information provided to the candidate.

Based on teacher response to the orientation materials for this assessment, we also suggest that such materials be revised to more clearly indicate the test content and objectives (perhaps by providing sample items), and that the materials should be kept as clear and concise as possible.

Assessment Content

Based on information collected from teachers and assessors, our analysis of the test items on two of the examination's six forms, and on performance results, we offer the following conclusions about the content of the Elementary Education Examination:

- Congruence of the test with the various California curriculum guides and frameworks is fair to high. Not all curricular emphases are reflected in the test items, however, and the balance of items across grade levels and subjects, especially in the area of science, could be improved.
- Generally, the breadth of coverage by this test of the California Standards for Beginning Teachers is good, but depth is lacking. The greatest need for improvement is in the areas of cross-cultural teaching and teaching students who are diverse with respect to needs, interests, and learning styles.
- The job-relatedness of this assessment was affirmed by a majority of the participating teachers. Those teachers who disagreed tended to be primary-grade teachers who perceived many of the test items as not being relevant to their grade level.
- The "materials-based items," which constituted a special feature of this assessment, were also judged by a majority of the participants to be

reflective of the tasks they perform as teachers. Some teachers, however, such as those of many Limited English Proficient students, felt the items reflected the "ideal" and not the "real" classroom.

- Although the majority of teachers did not think this test was too difficult for beginning teachers, the beginning teachers participating in this assessment answered, on average, only two-thirds to three-fourths of the items in each area correctly.
- An analysis of the content of two of the exam's six forms revealed a disproportionate number of items representing the subject areas and concepts associated with the upper grades (i.e., grades 4-8), as well as a lack of items addressing the teaching of diverse student groups (e.g., bilingual/LEP students, students characterized as low- or high-ability).
- The content analysis also indicated that the subsets of items measuring teaching competence varied across subject areas in the degree of sophistication required to answer questions, varying, for example, from the ability to match an activity with a particular approach to reading instruction to the ability to identify the best manipulative to teach a specific concept in math.
- Although the test was judged by teachers to be fair to new teachers of both genders, different ethnic groups, etc., 40% of the teachers did not think this assessment is an appropriate way to assess teacher competency. Many teachers objected to what they perceived as a focus on assessing content knowledge, and others on being tested on subject matter they have not taught. Several teachers suggested that the multiple-choice assessment be replaced by or supplemented with more direct measures of teaching such as interviews and classroom observations.

If the Elementary Education Examination is considered for further development, a content review by California educators should be conducted to examine each item for clarity, accuracy, sensitivity (or bias), job relevance, and relationship to the California curriculum guides and frameworks and credentialing requirements. They should also look at the test specifications for completeness, clarity, importance for credentialing, and job relevance. All items should also be reviewed by professional editors and sensitivity reviewers prior to field testing.

Assessment Format

The multiple-choice format will be discussed in more detail in Chapter 8 and contrasted with the classroom observation and semi-structured interview methods of teacher assessment. Its strengths appear to be ease and efficiency of administration and scoring, as well as an ability to cover a wide range of subject areas.

The format of this assessment could be improved in two respects. First, procedures to handle handicapped examinees, late arrivals, irregularities, or other special circumstances should be established. Second, the following problems with specific types of items were identified:

- The materials-based items were very lengthy, consisting of up to four pages of reference materials for one question. Since these items most directly reflect actual instructional decisions, they should not be abandoned, but the number of questions that are related to the most extensive reference materials should be increased.
- Items that require teachers to identify the "most" or "least" appropriate instructional technique constituted half of the items on the test. These questions were confusing, with reports of instances of marking the "most" appropriate answer when the "least" was required or vice versa. This format was especially common for items that assess pedagogical content knowledge. These items should be kept to a minimum, and consideration should be given to limiting the format to either "most" or "least" items throughout the test.

Two hours appeared to be a sufficient time in which to complete the test in its current format with about 77 items per form.

Summary

Compared to other multiple-choice examinations, the innovative elements of the Elementary Education Examination appear to increase the job relevance of a multiple-choice assessment. These innovative elements include the use of questions that explicitly relate to specific classroom contexts, and the use of teaching materials (such as student papers and teacher manuals) in the test booklet. However, taken as a whole, the examination needs revisions both in terms of improved balance across grade levels and subjects, and, within subjects, across the types of knowledge that are assessed (i.e., content knowledge, pedagogy, and pedagogical content knowledge). The relevance of this assessment approach to teachers who work with specific student populations also needs to be examined. Data were not available to evaluate the items separately by either format ("most/least" or materials-based items) or focus (content, general pedagogy, or content pedagogy).

CHAPTER 7:

SEMI-STRUCTURED INTERVIEW: SECONDARY SOCIAL SCIENCE

CHAPTER 7:

SEMI-STRUCTURED INTERVIEW: SECONDARY SOCIAL SCIENCE

The Semi-Structured Interview for Secondary Social Science (SSI-SSS) was developed by the Stanford Teacher Assessment Project (TAP). Like the SSI-EM, it was developed as a prototype of a type of examination to be used to certify distinguished master teachers. The Stanford TAP had previously administered the SSI-SSS to a sample consisting mostly of experienced teachers, but with a few student teachers and first-year teachers.

The SSI-SSS resembles the SSI-SM and SSI-EM in that it combines the Structured Interview and Assessment Center formats. It consists of three tasks:

(1) **Reviewing a Textbook:** A candidate reviews a textbook and completes a form which solicits a critique of specific aspects;

(2) **Planning a Lesson:** A candidate spends thirty minutes planning a lesson on a given topic and then responds to questions about that lesson; and

(3) **Use of Documents:** A candidate is given a group of documents to study, selects two as suitable for serving as the focal point of a series of lessons, and responds to questions about both their choice and the use of documents in social science classrooms.

Like the SSI-EM, the SSI-SSS required revision of the protocols based on the previous experience in administering the assessment. FWL and CTC/SDE staff met with one of the original test developers from the TAP, who provided guidance for changing the protocols. Due to previous time commitments, he could not make the changes himself. Assessors were recruited, and training, to be conducted by FWL staff, was scheduled for May 25.

In preparing for the assessor training, we became concerned about the feasibility of pilot testing the SSI-SSS for several reasons:

(1) The level of difficulty of the tasks was perceived to be high. We believed that teacher preparation programs do not instruct students in textbook review and the use of documents, and new teachers would be unprepared to do these tasks. Furthermore, the test developer had stated that he was very disappointed in the performance of master teachers in the first pilot test, which reinforced our reservations.

(2) Revision of the protocols included changes in the documents used. The test developer provided the new set of documents, while FWL staff made changes in the protocols. In reviewing the revised "Use of Documents" protocol, we were not certain that the questions were consistent with the revised set of documents. The test developer had

> no further time to revise the protocols before the scheduled training of assessors. We were uncomfortable about conducting the training of assessors when we ourselves did not understand the intent of some of the questions.

We reported these concerns to CTC/SDE staff, and it was decided to postpone the pilot test of the SSI-SSS until the protocols could be examined by experienced secondary social science teachers. After the experience with administering the SSI-EM, where the content was perceived to be too difficult for beginning teachers by both assessors and many teachers, it was decided to defer the pilot test of the SSI-SSS until 1989-1990. Another semi-structured interview in secondary social studies which has been specifically developed for beginning teachers by the State of Connecticut is also being considered for pilot testing.

CHAPTER 8:

CONCLUSIONS

CHAPTER 8:

CONCLUSIONS

This final chapter begins by summarizing our conclusions about the assessment approaches that were pilot tested during Spring 1989. We then suggest a framework for comparing our findings about the strengths and weaknesses of different assessment approaches. We conclude by identifying issues to be explored in the next round of developing and pilot testing assessment instruments, and decisions that should be made prior to selecting any assessment approach.

Assessment Approaches

Although the purpose of the pilot tests was to use the specific instruments to learn about the potential of assessment ***approaches***, the preceding chapters mainly focused on individual instruments. This section compares each instrument that was pilot tested to other instruments representing the same assessment approach, and summarizes our conclusions about the critical features as well as the strengths and weaknesses of each approach. These conclusions are based on our in-depth examination of one or two state-of-the-art instruments representing each approach. In formulating our conclusions in this section, we tried to go beyond our experience with each individual instrument to imagine the development of parallel indicators, tasks, or questions, either to extend the approach to new domains or to better address the domains of teacher competence that we examined.

Each instrument reflected one of three assessment approaches: classroom observation, semi-structured interview, or multiple-choice examination.

Classroom Observation

Definition. A classroom observation approach to teacher assessment consists of observing teachers as they instruct students in their classrooms to evaluate their performance. Two dimensions of classroom observation systems are: open vs. closed and low- vs. high-inference. In an open system, the observer attempts to describe "all" behaviors that occur without regard to selection or interpretation. In a closed system, the observer focuses on specific behaviors or categories of behavior. In open systems, evaluators judge the quality of performance without the benefit of a careful definition of what to look for in classrooms. Observation systems that are typically used for teacher assessment are closed systems. The low- vs. high-inference systems differ in the degree of specificity in the behaviors judged. In low-inference systems, criteria are defined in terms of specific behaviors, allowing little observer discretion. High-inference systems describe the behaviors more generally, requiring more use of the observer's judgment to identify and judge behaviors.

Characteristics of instruments piloted. As a state-of-the-art example of a classroom observation instrument, we pilot tested a high-inference classroom observation instrument, the Connecticut Competency Instrument (CCI). In addition, we examined written materials which described instruments used in Florida and Georgia to assess beginning teachers.

The CCI is a high-inference classroom observation system in which 10 indicators are grouped in three clusters to represent major aspects of instruction: management of the classroom environment, the instructional process and student assessment. Four assumptions underlie the design of the instrument and distinguish it from low-inference observation systems or competency check lists that are used as teacher credential requirements in other states: (1) it acknowledges that effective teachers practice in many different ways; (2) it focuses on general teaching abilities; (3) it is intended for beginning teachers; and (4) it emphasizes the importance of professional judgment in rating performances.

Strengths and weaknesses. Classroom observations assess teachers in the process of doing their work, so they have high job relevance and face validity. This assessment approach was usually cited by teachers who suggested a specific model when discussing the job relevance or appropriateness of the assessment in which they participated.

Susan Stodolsky (1989) has questioned the utility of classroom observations as a form of teacher assessment. Her research (1988) suggests that the subject being taught affects many teacher behaviors that are measured by common observation evaluation systems, such as the extent to which higher-order thinking skills are a focus of the lesson, the extent of student engagement, and the likelihood of observing teacher-directed (e.g., lecture) vs. student-centered (e.g., inquiry approaches) instructional formats. While many classroom observation systems that consist of behavioral check lists are vulnerable to this criticism, it does not seem to apply to the CCI, which relies on professional judgments that take into account the goal of the lesson as stated by the teacher and allows for differing approaches to the same goal.

However, Stodolsky also concludes (1988, p. 12) that "teaching is context dependent and stability is likely only within well-defined contexts, such as lesson types within subjects and grade levels." By allowing the teacher to select the lesson to be observed with no restrictions, a teacher can choose both the subject and the lesson approach that best displays his/her competence. With a limited number of observations, it is impossible to obtain a complete sample of teacher behaviors across all subjects, grade levels, and lesson approaches taught during a year. This sampling problem is not, of course, peculiar to classroom observations, but is true of other assessment approaches as well.

Classroom observations seem to be best suited to assess the actual application of general principles of teaching in the classroom. Significant types of teacher classroom performance can be carefully defined, and the methodology is available to overcome technical problems of inconsistent observations. However, not all knowledge domains can be seen in classroom behaviors. Compared with other approaches to assessment, classroom observations seem to be most appropriate for assessing limited samples of content knowledge, and broader samples of general pedagogy, content pedagogy, knowledge of learners and learning, and management of classroom climate.

Semi-Structured Interviews

Definition. Semi-structured interviews provide opportunities for candidates to respond orally to a standardized series of questions or tasks that are presented verbally by an examiner who uses a script known as an interview schedule. Semi-structured interviews include "probes" to be used at the administrator's discretion to enable candidates to elaborate on their responses.

Characteristics of instruments piloted. Both the SSI-SM and the SSI-EM, which serve as state-of-the-art exemplars of semi-structured interviews, draw largely from Gaea Leinhardt's research in mathematics in which teachers perform tasks and then explain their responses. We found few examples of interview assessments, so the SSI-SM and SSI-EM serve as early prototypes of the interview approach to the evaluation of teachers.

Strengths and weaknesses. The strength of the semi-structured interview is its ability to assess depth of knowledge, especially knowledge that impacts instructional planning and decision-making. The semi-structured interviews have a good potential to adequately sample knowledge domains with a limited range, (e.g., educational philosophies and goals). Therefore, we see semi-structured interviews as the best approach to assess knowledge of learners and learning, the effects of school context, and educational philosophies and goals. For other domains with a broad range, such as curriculum knowledge, general pedagogy, and content pedagogy, semi-structured interviews assess the subject, topic and grade level addressed in the interview quite well, but the coverage of these domains is only partial. The direct relationship of teacher responses during interviews to their actual classroom behaviors is limited (except for content knowledge).

To our knowledge, the question of whether teachers are able to implement the activities described during interviews has not been systematically studied. One researcher (Wilson, 1988) observed classrooms and discussed teacher patterns of behavior with supervisors of a small group of teachers participating in semi-structured interviews. She found that the activities described during the interviews were consistent with those observed or described by supervisors; however, this study was in conjunction with voluntary participation in a research project. In the case of a credentialing requirement, teachers would have a much greater incentive to represent as strong a performance as possible in the interview.

It is critical for semi-structured interviews to contain questions that explicitly address all the areas to be scored. The degree of explicitness needed was underestimated by the developers of the two semi-structured interview prototypes which were pilot tested. Without either explicit questions or the ability to probe, it is impossible to know whether a teacher's lack of responses in a particular area are due to a lack of clarity or a lack of ability. Moreover, the use of discretionary probes in a "high stakes assessment" could serve to prompt correct responses among some of the teachers being assessed. The dynamics of probing ambiguous responses should be studied to see if probes are either necessary or desirable as part of the semi-structured interview approach to teacher assessment.

Like observations, well-designed interviews can meet standards of technical quality. Unanswered yet is the question of the amount of distinctive information that interviews add to assessment decisions. Studies of convergent and discriminant validity are needed before any combination of measurement approaches considered in a teacher evaluation system.

Multiple-Choice Examinations

Definition. Multiple-choice examinations are those in which each candidate selects a correct response from a fixed number of response options. Scoring is typically on a right-wrong basis for each item, though other scoring systems that grant partial credit or deduct for guessing are also used.

Characteristics of instruments piloted. The major difference between the Elementary Education Examination and other multiple-choice teacher assessments (such as the NTE) is the attempt to embed both theoretical and applied questions (e.g., implications of Piaget's theory) in classroom contexts. Another important difference is the inclusion of materials-based items that require candidates to read and evaluate documents such as Individual Education Plans, student worksheets, and report cards. These innovations were designed to make the tasks on the test more similar to those in the classroom. Such "innovative" items were developed for each subject domain, suggesting that it is possible to create multiple-choice items to assess content pedagogy which use common teaching materials. When teachers were asked to comment specifically on these innovative items, about two-thirds felt that they ***did*** reflect actual teaching tasks.

The innovative items were most often presented in an undefined classroom context, implying that the appropriateness of classroom activities is unaffected by the type of students taught. Some teachers questioned the relevance of such general items to their competence in teaching the particular type of students (generally either low-achieving or Limited English Proficient) who were most prevalent in their classroom.

Strengths and weaknesses. The strength of multiple-choice examinations is their ability to sample a broad range of aspects of a knowledge domain, as long as they are represented in a fixed-response format. The "one right answer" format is a major limitation for evaluating many teaching tasks where multiple approaches are both appropriate and desirable, especially where the appropriateness is at least partially dependent on the context (i.e., students, subject, grade-level, and teacher characteristics). The depth with which multiple-choice items assess this knowledge is limited and the relationship between performance on the examination and actual classroom practice is modest.

Framework for Comparing Differing Assessment Approaches

The measures of teacher performance that were pilot tested in 1989 represent three quite different assessment approaches: observations, multiple-choice examinations, and interviews. Each of these approaches allows direct, authentic measurement of some domains of teacher performance, but is less suitable for providing information about other domains of performance. To compare the strengths and weaknesses of different assessment approaches, it is important to have a framework which includes a broad range of domains of teaching competencies. In this section, we describe such a framework, but we propose that it be used after a greater number of assessment approaches have been pilot tested.

To identify a broad range of domains of teacher performance, we began with a review of literature on teacher assessment that discussed teacher competencies, or the knowledge base of teaching. We found surprisingly few publications that explicitly delineated competencies (See Shulman, 1987; Wilson, 1988; Leinhardt, 1989). We also drew from our observations of the assessments that we pilot tested and of two other assessments pilot tested by the Stanford Teacher Assessment Project in the summer of 1989. After this review, we decided that a list compiled by Shulman (1987) seemed to be a good summary. To Shulman's list, we added two additional domains, Classroom Climate and Professional Collegiality. We do not advocate that beginning teacher knowledge and/or practice be assessed in all these domains, but we believe that it is important to identify as complete a range as possible to fully compare different assessment approaches.

The domains identified are:

(1) **Content Knowledge:** Understanding principles and concepts in the subject(s) taught, their interrelationships, and their application to other related content areas.

(2) **Knowledge of Curriculum:** Understanding hierarchical and nonhierarchical relationships between the concepts and principles of the content area which guide construction of curriculum, resources available to teachers, and of theories guiding the development of curriculum.

(3) **General Pedagogy:** Understanding and using generalizable approaches and methods for managing, planning, presenting and assessing in struction.

(4) **Content Pedagogy:** Understanding alternative representations of the subject matter, knowledge of student conceptions related to particular topics, and pedagogical reasoning related to the particular content being taught.

(5) **Knowledge of Learners and Learning:** Understanding how to tailor instruction to student culture, language, interests, background experiences, and cognitive and physical abilities.

(6) **Management of Classroom Climate:** Understanding approaches for creating an optimal physical and psychological environment for learning, maintaining rapport with students, setting high expectations for achievement, and establishing norms for student-student interactions.

(7) **Knowledge of Effects of School Context:** Knowledge of the organizational, political, cultural and social context of the school.

(8) **Knowledge of Educational Philosophies, Goals, and Objectives:** Comparative knowledge of educational philosophies, goals, and objectives, including their bases and justifications.

(9) **Professional Collegiality:** Knowledge and disposition for continuing professional development and collaboration with colleagues.

Additionally, to evaluate how well an approach can measure each of the nine domains, we recommend that California consider the extent to which each assessment could assess three dimensions of each knowledge domain:

- **Sampling:** the number and range of aspects (e.g., concepts, contexts, situations, skills) that the assessment approach can tap. The key issue is how broadly the modality can sample the domain of knowledge/skills.

- **Depth:** the extent to which the focal skills, tasks, questions, or responses provide evidence of the teacher's knowledge, understanding, or reflective reasoning about the domain.

- **Application:** the degree to which the focal skills, tasks, questions, or responses match the teachers' thoughts and actions as they occur in actual teaching situations.

Using this framework, hybrid forms of assessment approaches can be designed to take advantage of strengths and to compensate for the weaknesses of a single approach. Examples include: an observation and a semi-structured interview; an observation and written responses; a semi-structured interview with a portfolio. We anticipate using this framework for the comparison of assessment approaches, including those pilot tested this year, at the end of the next phase of pilot testing.

Table 8.1 illustrates how evaluations or ratings of assessment approaches might be summarized across the nine teaching domains and three evaluation dimensions, using three hypothetical assessment approaches. The table would be interpreted in the following way: The strengths of Approach 1 lie in its ability to assess a teacher's ability to apply their knowledge in an actual teaching situation, especially in certain domains; the major limitations is that it does not do so in depth. The potential for sampling entire domains is more limited; the best it achieves is a partial sampling of a few domains. By contrast, Approach 2 can assess a teacher's knowledge in depth in most areas, but exhibits a much weaker ability to assess a teacher's ability to apply knowledge. Its sampling ability is stronger than that of Approach 1 for most domains. The third approach is the strongest of all three in its sampling ability, which is its strength; it is weak in its ability to assess a candidate's knowledge either in depth or in relation to classroom application.

Cost Estimates

Assessments that strive to validly evaluate teaching competence and performance are more expensive than traditional multiple-choice examinations. Given the developmental nature of the assessment instruments that were pilot tested, the cost estimates in this report were included to illustrate the ingredients which compose the various assessment approaches, and not to provide accurate cost estimates to be used in policy decisions. Actual cost estimates are highly sensitive to decisions regarding the administration of an assessment. For example, Connecticut reduces the costs of administering the CCI by training large numbers of observers who are then provided up to six days of release time by their employers, with the Connecticut State Department of Education responsi-

TABLE 8.1

ANALYSIS OF ALTERNATIVE ASSESSMENT APPROACHES AND THEIR ABILITY TO ASSESS SPECIFIC TEACHING COMPETENCIES

PERFORMANCE DOMAIN	ASSESSMENT APPROACHES								
	APPROACH 1			APPROACH 2			APPROACH 3		
	Sampling	Depth	Application	Sampling	Depth	Application	Sampling	Depth	Application
Content Knowledge	⊙	○	●	⊙	●	◒	●	⊙	⊙
Knowledge of Curriculum	⊙	○	⊙	◒	●	⊙	⊙	○	⊙
General Pedagogy	◒	○	●	◒	●	○	●	⊙	○
Content Pedagogy	◒	○	●	◒	●	⊙	●	⊙	○
Knowledge of Learners and Learning	◒	○	◒	●	●	⊙	●	⊙	○
Management of Classroom Climate	◒	○	●	⊙	⊙	○	⊙	⊙	○
Knowledge of Effects of School Context	⊙	⊙	⊙	●	●	○	●	⊙	○
Knowledge of Educational Philosophies and Goals	○	○	○	●	●	○	●	○	○
Professional Collegiality	○	○	○	⊙	⊙	○	⊙	⊙	○

Extent of Coverage

●	Extensive
◒	Partial
⊙	Very limited
○	Not at all

ble for only travel expenses and the cost of any substitute teachers. When California is ready to identify the most suitable approach (or combination of approaches) to assessment, an important step will be to design a cost-effective method of administration that preserves the technical quality of the approach.

Next Steps in Designing a System of Teacher Assessments

The construction of a system of assessment for teachers has two major components: the selection of assessment instruments and the design of an assessment system using these instruments. In this section, ways in which future pilot tests for the California New Teacher Project can be used to inform choices of assessment instruments are discussed, then some major decisions to be made in connection with the design of an assessment system are identified.

Future Pilot Tests

The final two years of the three-year California New Teacher Project include plans for the development and pilot testing of assessments which exemplify approaches other than those discussed in this report. Assessment approaches to be pilot tested in the spring of 1990 include: constructed written responses, structured simulations, analysis of videotapes, and portfolio review. The subject areas to be assessed will include secondary English, secondary science, and elementary teaching. In addition, the Interagency Task Force plans to commission the development of additional assessments as needed for the pilot study. These two years of pilot testing provide opportunities for exploring issues related to evaluating assessments. Given budgetary and time constraints, issues must be carefully chosen, since not all can be explored. The issues discussed below are only a representative rather than an exhaustive list, but include what we believe to be major questions unanswered by this initial round of pilot testing in 1989.

- **Development of scoring.** In the development of instruments, it is important that issues of scoring be considered at an early stage. No instrument should be pilot tested until it has been subjected to a small-scale administration (with as few as two or three teachers) to see that the stimulus materials, questions, or tasks are eliciting scorable responses. Once this stage is accomplished, the larger pilot study can concentrate on ascertaining how the various subparts form either a single construct or separate factors, how they correlate between raters, and how to set passing scores both at the initial and later stages of administration.

- **Commissioning assessments in different subjects.** The piloting of different assessment approaches and assessments that measure the ability to teach different subjects provides an opportunity not only to evaluate the ability of assessment approaches to measure teaching skills in a variety of subjects but also to deepen our knowledge about teaching skills. Estimates of the ability of specific assessment instruments and approaches to evaluate teaching skills in a variety of subjects can be facilitated by explicitly identifying similarities and differences between teaching skills in various subjects. For example, noncognitive skills including esthetic, aural, and physical abilities play a greater role in teaching the arts, foreign language, and physical education than in mathematics, social studies, reading, English, or science. It is likely that teachers of the former set of subjects need some skills that are not required by teachers of the latter set of subjects. The development of an assessment of middle school teaching could also assist in the identification of differences in teaching skills required at various grade levels. Additional skills identified in the research literature can guide assessment developers as they construct the assessments; it is possible that other skills will be identified during the course of the development and analysis of specific assessment instruments.

- **Greater range of teaching experience in sample.** Evaluating the appropriateness of particular assessment instruments and assessment approaches for beginning teachers can be informed by including student teachers, beginning teachers, and more experienced teachers in the sample of teachers participating in the pilot tests. Systematic group differences in performance can facilitate the identification of stages in the development of specific competencies, and can guide the choice of both assessment approach and the time when it is administered.

- **Multiple assessment of teachers.** Although we relied on our experience from the administration and analysis of assessment instruments as a basis for evaluating assessment approaches, assessing the same teachers with different instruments would provide more explicit data, especially for the comparison of assessment approaches. For instance, teachers could be observed with the CCI and participate in an additional assessment addressing the subject taught in the lesson observed. This design would increase our ability to assess the extent to which different assessments provide supplementary and complementary information. We could identify tradeoffs when one assessment instrument is chosen over another, and redundancies if more than one is used. In the absence of a general measure of teaching ability, these comparisons would also contribute information about the validity of the measurement of teaching skills that are assessed by more than one instrument.

- **Ability to teach diverse students.** In 2989 we found that no instrument provided a good model for assessing competencies related to the teaching of diverse students. Assessment instruments that have been commissioned by the California New Teacher Project for Spring 1990 pilot testing include an emphasis on the teaching of diverse students, an issue of increasing importance to California educators. The evaluation of the success of assessment instruments in addressing this issue can guide the identification of the next steps to take and pitfalls to avoid in building this component into all assessments.

- **Content review.** Prior to adoption, all assessment instruments should undergo review of teaching skills and subject matter content by California teachers, teacher educators, subject matter specialists, and experts in performance assessment. Although all instruments commissioned by the California New Teacher Project have included some review by groups of California educators, a wider review is needed to assess congruence with current research and professional norms.

Issues in Design of an Assessment System

As the previous discussion in this chapter has indicated, the utility of assessment approaches and instruments cannot be evaluated apart from their purpose. The design of an assessment system includes prerequisite decisions about the purpose of assessment, which then guide the selection of assessment instruments. We end this chapter by identifying what we see as the ***major decisions*** guiding the selection of instruments for use in a system of assessment of teachers.

- **Assessment focus.** Perhaps the most crucial issue is to decide what competencies are to be used as screens at particular stages of teacher preparation and teaching. Currently, prospective California teachers are required to demonstrate basic reading, writing and mathematical skills and subject matter competence to obtain a teaching credential. Decisions about requiring passage of some type of performance assessment will depend upon the relative priorities assigned to specific teaching skills, as well as on assumptions about the degree to which these skills are likely to be developed in the beginning years of teaching.

- **Breadth of assessment.** Another issue concerns how to address the multiple grades, subjects/topics, and contexts which are covered by a specific teaching credential. The multiple subjects credential perhaps represents the most difficult case, because it covers the broadest range of subjects and grade levels. To what degree should a range of grades, subjects/topics and contexts be sampled to provide sufficient assurance of the competence of the teacher candidate to instruct effectively in all areas covered by the credential? How should a candidate's experience teaching at specific topics to specific groups of students be utilized? What should be the relative degree of emphasis between breadth of sampling and the depth of knowledge gained from experience?

- **Flexibility of Assessment System.** Since a multi-stage, multi-year credentialing system is envisioned, and since different teachers develop at different speeds, the extent of flexibility of an assessment system should also be considered. Should only candidates with scores near the passing threshold have to take more complex tests? Could candidates be allowed to compensate for weaknesses in some areas through strengths in other areas, providing that some minimal competence has been demonstrated in the weaker areas? If so, what would be required of ***all*** candidates and where should such minimal thresholds be placed?

- **Coordination with professional development.** To what extent should credentialing decisions and professional development be coordinated? The most expensive assessments also tend to provide the best guidance for training activities to improve teaching. Several states provide professional development for candidates who fail their first performance assessment, providing they are rehired by their employing district. The use of information from credentialing assessments for staff development might justify the greater expense of those assessments which provide more information.

- **Interagency involvement.** Another decision concerns the role of state, regional, and local agencies in the credentialing process. The administration of assessments could range from a highly centralized program with professional assessors selected and trained by the Commission on Teacher Credentialing to a program resembling the current system of credentialing teachers, in which assessments are regionally administered by institutions of higher education following guidelines established by the Commission. In Connecticut, teachers, teacher educators, and state agency staff who are trained by the state and provided release time by their employing organizations to administer the CCI. Other variations are possible. One example is that the CTC could develop and oversee the implementation of an assessment system through a decentralized network of agencies, including universities, county offices of education, and school districts. Options such as these need to be outlined, broadly discussed, and used to evaluate the various assessment options and designs.

- **Funding.** Another significant decision to be discussed concerns funding. Given that teaching is a relatively low-paying profession and that an increase in the supply of teachers is needed for the foreseeable future, it is unlikely that new teachers could be expected to bear the full cost of the new assessment approaches. Even apart from fees charged, new teachers would incur opportunity costs in terms of time spent preparing for and participating in the new assessments. If assessment approaches which lend themselves to centralized administration were chosen, then rural teachers who live far from assessment sites would not only spend greater amounts of time traveling to the assessment site, but would also incur additional expenses for travel and perhaps overnight lodging.

While there is a great potential for performance-based assessments to highlight and strengthen the knowledge and skills that teachers should possess, it is not clear how to balance increasing standards and the increasing needs for teachers from diverse backgrounds who might be the least able to afford more costly assessments. Options need to be outlined which will provide alternative ways of balancing these concerns.

The choice of assessment approaches and the design of a system of teacher assessment are conditioned by a series of decisions concerning the purpose of the assessments. For this reason, we have not identified the one best assessment approach or the best design for an assessment system. Instead, we have summarized what was learned from the pilot testing of three specific assessment approaches, suggested issues that could be explored with the next round of developing and pilot testing assessment instruments, and outlined issues which should be considered before selecting one or more assessment approaches for use in credentialing decisions.

We see a great opportunity during the next two years to use activities planned for the CNTP to identify knowledge, options, benefits and costs for strengthening teacher assessment and credentialing. First, the alternative approaches that have been commissioned by the state agencies for piloting in the second year of the project give promise of high validity with respect to representing authentic teaching behaviors. Second, the pilot testing of a variety of assessment approaches can contribute to the identification and discussion about the various ways in which these assessments might be used such that they: (1) reflect state-of-the-art knowledge in the areas of curriculum, pedagogy, and the teaching of diverse student populations; (2) support and direct teacher preparation and staff development programs by highlighting important and critical knowledge and skills; (3) reflect the complexity of teaching, thus increasing its attractiveness and professionalization; and (4) help attract, rather than discourage, the strongest and most diverse teacher candidates. Finally, information from the pilot tests can help develop alternative funding mechanisms that will not be unduly burdensome to teachers or local or state agencies.

BIBLIOGRAPHY

BIBLIOGRAPHY

Berliner, David. (1989). *Implications of Studies of Expertise in Pedagogy for Teacher Education & Evaluation* in ***New Directions for Teacher Assessment,*** invitational conference proceedings. Princeton, NJ: Educational Testing Service.

Berliner, BethAnn, Mata, Susana, Zalles, Dan, Little, Judith Warren. (1987). *Improving student teaching through clinical supervision.* ***Volume Two: Supervision and support through the eyes of student teachers and first year teachers.*** San Francisco, CA: Far West Laboratory for Educational Research and Development.

Berliner, BethAnn, Intili, JoAnn, Little, Judith Warren, Mata, Susana, Terry, Patricia, Zalles, Dan. (1987). *Preserving preservice teacher education through clinincal supervision of student teaching.* ***Volume four: The university supervisor: Program impact and experience.*** San Francisco, CA: Far West Laboratory for Educational Research and Development.

Borko, H. (1986). *Clinical teacher education: The induction years. In James Hoffman and Sara Edwards, eds.,* ***Reality and reform in clinical teacher education, (pp. 45-64).*** New York, NY: Random House.

Borko, Hilda, Lalik, Rosary, Livingston, Carol, Pecic, Kathleen, and Perry, Diana. (1986). ***Learning to teach in the induction year: Two case studies.*** Paper presented at the annual meeting of the American Educational Research Association.

Boyer, Ernest L. (1983). ***High school: A report to the Carnegie Foundation for the advancement of teaching.*** New York: Harper & Row.

California State Department of Education. (1985). ***Mathematics Framework for California Public Schools, Kindergarten through Grade Twelve.*** Sacramento: California State Department of Education.

California State Department of Education. (1987). ***Science-Model Curriculum Guide, Kindergarten through Grade Eight.*** Sacramento: California State Department of Education.

California State Department of Education. (1988). ***English-Language Arts, Model Curriculum Guide, Kindergarten through Grade Eight.*** Sacramento: California State Department of Education.

California State Department of Education. (1988). ***History-Social Science Framework, Kindergarten through Grade Twelve.*** Sacramento: California State Department of Education.

Clark, D.C., Smith, R.B., Newby, T.J., & Cook, V.A. (1985). *Perceived origins of teach ing behavior.* ***Journal of Teacher Education, 36(6), 49-53.***

Gomez, Robert. (1989). ***A report on teacher supply: Enrollments in professional preparation programs in California institutions.*** Sacramento: Commission on Teacher Credentialing.

Goodlad, John I. (1984). ***A place called school: Prospects for the future.*** New York: McGraw-Hill.

Grant, Carl and Zeichner, Kenneth. (1981). *Inservice support for first year teachers: The State of the scene,* ***Journal of Research and Development in Education, 14:99-111.***

Holmes Group, Inc. (1986). ***Tomorrow's teachers: A report of The Holmes Group.*** East Lansing, MI: The Holmes Group.

Huling-Austin, Leslie. (1988). ***A synthesis of research on teacher induction programs and practices.*** Paper presented at the annual meeting of the American Educational Research Association.

Leinhardt, Gaea. (1989). *Math Lessons: A Contrast of Novice and Expert Competence.* ***Journal for Research in Mathematics Education, 20, 52-75.***

Lortie, Dan. (1975). ***Schoolteacher.*** Chicago: University of Chicago Press.

Mata, Susana, Berliner, BethAnn, Intili, JoAnn, Little, Judith Warren, Stansbury, Kendyll. (1988). ***Final report. Improvement of preservice teacher education through clinical supervision of student teachers.*** San Francisco, CA: Far West Laboratory for Educational Research and Development.

McDonald, F.J. (1980). *Study of induction programs for beginning teachers.* ***Volume one: The problems of beginning teachers: A crisis in training.*** Princeton, NJ: Educational Testing Service.

National Commission on Excellence in Education. (1983). ***A nation at risk: The imperative for educational reform.***

Odell, Sandra. (1986). *Induction support of new teachers: A functional approach.* ***Journal of Teacher Education. 26-29.***

Shulman, L.S., and Sykes, G. (1986). ***A national board for teaching? In search of a bold standard.*** A report for the Task Force on Teaching as a Profession. New York: Carnegie Corporation.

Shulman, Lee. (1987). Knowledge and teaching: ***Foundations of the New Reform.*** Harvard Educational Review, 57, 1-22.

Streifer, P. (1984). ***The Validation of Teaching Competencies in Connecticut.*** (An unpublished Ph.D. thesis).

Varah, Leonard, Theune, Warren and Parker, Linda. (1986). *Beginning teachers: Sink or swim?* ***Journal of Teacher Education, 30-34.***

Veenman, Simon. (1984). *Perceived problems of beginning teachers.* ***Review of Educational Research, 54, 143-178***

Watkins, Richard. (1985). ***A practitioner review of the content validity and passing standards of the California Basic Educational Skills Test.*** Sacramento: Commission on Teacher Credentialing.

Wilson, Suzanne. (1988). ***Understanding Historical Understanding: Subject matter knowledge and the teaching of Teachers.*** A dissertation submitted to Stanford University.

Wheeler, Pat, Hirabayashi, J.B., Maretinson, J., and Watkins, R.W. (1988). ***A study on the appropriateness of fifteen NTE specialty area tests for use in credentialing in the state of California.*** Emeryville, CA: Educational Testing Service.

Wheeler, Pat, and Elias, P. (1983). ***California Basic Educational Skills Test: Field test and validity study report.*** Berkeley, CA: Educational Testing Service.

Wise, Arthur, Darling-Hammond, Linda, Berry, Barnett, Klein, Stephen P. (1987). ***Licensing teachers: Design for a teacher profession.*** Santa Monica, CA: The Rand Corporation.

APPENDIX A:

SSI-SM RELIABILITY

APPENDIX A: SSI-SM RELIABILITY

Two types of reliability were examined: inter-rater consistency and internal consistency. Analyses were conducted on a subset of the assessment sample consisting of the ten teachers whose data were available. Due to this very small sample size, these analyses can be interpreted as exploratory investigations only.

Item Aggregation

Six groups of tests were created. Each group was based on a different level of aggregation of the 20 (two topics by two tasks by five indicators) items. Group 1 (Total Test) incorporated all 20 items into one test. Group 2 (Topic) separately incorporated the ten task-by-indicator items for each topic into two task level tests. Similarly, Group 3 (Task) separately combined the ten topic-by-indicator items for each task into two task level tests. Group 4 (Topic by Task) incorporated the five indicator items into four topic-by-task level tests. Group 5 (Indicators) combined the four topic-by-task items into five indicator level tests. Finally, Group 6 (Topic by Task by Indicators) combined each pair of topic items into ten task-by-indicator level tests.

Inter-rater Consistency

The primary analyses for evaluation of the inter-rater consistency of the tests are presented in Table 1. These consisted of calculating item means for each test separately for the first and second raters on that test (R1, R2) and for the averages from the two sets of ratings (RS). Additionally, inter-rater correlation coefficients between raters are presented for each test. The differences between item means across raters are uniformly small for all levels of test aggregation, indicating that different raters were employing similar standards within items for all item aggregations. The inter-rater correlations are mostly moderate. The correlation for the full test is .60. For other tests the correlations cluster around this level, with those tests having fewer items showing more variation among correlations.

As part of the pilot assessment administration, all pairs of raters were instructed to discuss their ratings for each indicator after the initial rating was made and try to reach a consensus. Whether or not they were able to reach a consensus, they then each made the rating again after considering any new information derived from the discussion. This process resulted in a duplicate set of ratings for all teachers. The analyses were repeated on these consensus ratings, and are summarized in Table 2. Here the small between-rater differences become even smaller and in most cases disappear, and the inter-rater correlations, with one exception, approach unity for all tests. These results indicate that the raters were able to reach or move toward consensus in most instances as a result of their discussions.

Because each task employed a different pair of raters, it was possible to explore the effect of reversing rater pairs in the construction of those tests which aggregated across tasks (groups 1, 2, and 5). As shown in Table 3, these reversals had little effect on item means, but substantially increased the inter-rater correlations. This

TABLE 1

MEAN ITEM RATINGS, COEFFICIENT ALPHAS, AND INTER-RATER CORRELATIONS FOR TESTS BASED ON INDICATORS

TEST	Ni	MEAN ITEM RATING R1	R2	RS	COEFFICIENT ALPHA R1	R2	RS	INTR RATR CORR
T1 (All items)	20	1.9	1.8	1.8	.85	.88	.87	.60
T2A (Topic 1)	10	2.1	2.0	2.0	.89	.79	.88	.56
T2B (Topic 2)	10	1.7	1.7	1.7	.72	.79	.77	.70
T3A (Task 1)	10	1.8	1.7	1.7	.90	.92	.92	.66
T3B (Task 3)	10	2.0	2.0	2.0	.59	.83	.77	.59
T4A (Tpc 1, Tsk 1)	5	1.9	1.6	1.7	.95	.91	.95	.57
T4B (Tpc 1, Tsk 3)	5	2.2	2.3	2.3	.89	.72	.84	.33
T4C (Tpc 2, Tsk 1)	5	1.7	1.7	1.7	.86	.81	.87	.70
T4D (Tpc 2, Tsk 3)	5	1.7	1.6	1.7	.87	.83	.87	.76
T5A (Indicator 1)	4	2.0	2.0	2.0	.25	.69	.59	.65
T5B (Indicator 2)	4	2.0	2.0	2.0	.38	.51	.41	.56
T5C (Indicator 3)	4	2.1	2.0	2.1	.25	.20	.19	.39
T5D (Indicator 5)	4	1.5	1.5	1.5	.55	.64	.66	.69
T5E (Indicator 6)	4	1.7	1.7	1.7	.31	.78	.54	.32
T6A (Tsk 1, Ind 1)	2	1.8	1.8	1.8	.17	.88	.86	.63
T6B (Tsk 1, Ind 2)	2	2.2	2.0	2.1	.45	.54	.42	.54
T6C (Tsk 1, Ind 3)	2	2.0	1.8	1.9	.32	.86	.62	.49
T6D (Tsk 1, Ind 5)	2	1.6	1.5	1.6	.75	.70	.80	.88
T6E (Tsk 1, Ind 1)	2	1.4	1.4	1.4	.64	.71	.66	.51
T6F (Tsk 3, Ind 6)	2	2.2	2.2	2.2	.64	.73	.36	.74
T6G (Tsk 3, Ind 2)	2	1.9	1.9	1.9	.11	.62	.38	.63
T6H (Tsk 3, Ind 3)	2	2.3	2.3	2.3	.48	.28	0	.48
T6I (Tsk 3, Ind 5)	2	1.5	1.5	1.5	.16	.60	.34	0
T6J (Tsk 3, Ind 6)	2	2.0	2.1	2.0	0	.71	0	.12

TABLE 2

MEAN ITEM RATINGS, COEFFICIENT ALPHAS, AND INTER-RATER CORRELATIONS FOR TESTS BASED ON CONSENSUS INDICATORS

		MEAN ITEM RATING			COEFFICIENT ALPHA			
TEST	Ni	R1	R2	RS	R1	R2	RS	INTR RATR CORR
T1 (All items)	20	1.9	1.8	1.9	.85	.86	.86	.99
T2A (Topic 1)	10	2.0	2.0	2.0	.85	.86	.86	.98
T2B (Topic 2)	10	1.7	1.7	1.7	.74	.75	.75	1.00
T3A (Task 1)	10	1.7	1.7	1.7	.91	.90	.91	1.00
T3B (Task 3)	10	2.0	2.0	2.0	.68	.78	.75	.99
T4A (Tpc 1, Tsk 1)	5	1.8	1.8	1.8	.90	.90	.90	1.00
T4B (Tpc 1, Tsk 3)	5	2.3	2.3	2.3	.81	.82	.83	.95
T4C (Tpc 2, Tsk 1)	5	1.6	1.6	1.6	.84	.85	.85	.99
T4D (Tpc 2, Tsk 3)	5	1.7	1.7	1.7	.81	.82	.82	1.00
T5A (Consensus 1)	4	2.0	2.0	2.0	.33	.38	.36	.98
T5B (Consensus 2)	4	2.0	2.0	2.0	.38	.37	.38	.99
T5C (Consensus 3)	4	2.0	2.0	2.0	.37	.37	.37	1.00
T5D (Consensus 5)	4	1.6	1.5	1.5	.52	.42	.48	.99
T5E (Consensus 6)	4	1.7	1.7	1.7	.30	.62	.54	.88
T6A (Tsk 1, Con 1)	2	1.7	1.7	1.7	.58	.58	.58	1.00
T6B (Tsk 1, Con 2)	2	2.2	2.1	2.1	.40	.42	.42	.95
T6C (Tsk 1, Con 3)	2	1.8	1.8	1.8	.77	.77	.77	1.00
T6D (Tsk 1, Con 5)	2	1.5	1.5	1.5	.77	.63	.72	.99
T6E (Tsk 1, Con 6)	2	1.4	1.4	1.4	.86	.86	.86	1.00
T6F (Tsk 3, Con 1)	2	2.3	2.2	2.2	.44	.49	.47	.97
T6G (Tsk 3, Con 2)	2	1.9	1.9	1.9	.27	.27	.27	1.00
T6H (Tsk 3, Con 3)	2	2.3	2.3	2.3	.43	.43	.43	1.00
T6I (Tsk 3, Con 5)	2	1.6	1.6	1.6	.53	.53	.53	1.00
T6J (Tsk 3, Con 6)	2	2.0	2.1	2.0	0	0	0	.27

TABLE 3

MEAN ITEM RATINGS, COEFFICIENT ALPHAS, AND INTER-RATER CORRELATIONS FOR TESTS BASED ON INDICATORS, RATER COMBINATIONS REVERSED

		MEAN ITEM RATING		COEFFICIENT ALPHA		INTR RATR COR
TEST	Ni	R12	R21	R12	R21	
T1 (All items)	20	1.9	1.8	.85	.82	.80
T2A (Topic 1)	10	2.1	1.9	.89	.76	.62
T2B (Topic 2)	10	1.7	1.7	.67	.81	.72
T3A (Task 1)	10					
T3B (Task 3)	10					
T4A (Tpc 1, Tsk 1)	5					
T4B (Tpc 1, Tsk 3)	5					
T4C (Tpc 2, Tsk 1)	5					
T4D (Tpc 2, Tsk 3)	5					
T5A (Indicator 1)	4	2.0	2.0	.39	.39	.84
T5B (Indicator 2)	4	2.0	1.9	.36	.40	.69
T5C (Indicator 3)	4	2.1	2.0	.18	0	.57
T5D (Indicator 5)	4	1.6	1.5	.71	.31	.76
T5E (Indicator 6)	4	1.7	1.7	.63	0	.80
T6A (Tsk 1, Ind 1)	2					
T6B (Tsk 1, Ind 2)	2					
T6C (Tsk 1, Ind 3)	2					
T6D (Tsk 1, Ind 5)	2					
T6E (Tsk 1, Ind 6)	2					
T6F (Tsk 3, Ind 1)	2					
T6G (Tsk 3, Ind 2)	2					
T6H (Tsk 3, Ind 3)	2					
T6I (Tsk 3, Ind 5)	2					
T6J (Tsk 3, Ind 6)	2					

suggests that the level of agreement between rater pairs is partially dependent upon the individual raters employed. With a larger sample size, however, it is possible that inter-rater correlations would be more similar between different rater pairs.

Internal Consistency

Coefficient alphas also are presented in Table 1. These are similarly moderate to high, mostly from the high seventies to the mid-nineties, for aggregations across all items; across tasks and indicators; across topics and indicators; and across indicators. For aggregations across topic and task and across topic only, where fewer items are available to aggregate, the alphas are much smaller and less stable, with many falling below .5. These results might suggest that neither topic/task nor topic are not appropriate levels for aggregation, or alternatively, could result primarily from the small n's. To investigate this second possibility we used the Spearman Brown method to estimate alphas for tests of equal test length (n=10) at all aggregations. These are presented in Table 4. Here the alphas for the task/level aggregation (group 5) still are unacceptably low, but those for the topic level aggregation (group 6) appear to have risen to sufficient levels. Caution is required, however, in interpreting results of estimations for longer tests based on two item tests as it is unlikely that eight additional items would be sufficiently similar to the first two to meet the assumptions of the adjustment procedure. This, combined with the small sample, make these results very exploratory. With that warning in mind, it seems that these data justify aggregations across all items; across tasks and indicators; across topics and indicators; and across indicators. Aggregation across topics and tasks is contraindicated, while aggregation across topics alone is marginally supported.

Table 2 shows that consensus ratings had little effect on internal consistency for most tests; while Table 3 similarly indicates little effect on the alphas of reversing rater pairs.

Dichotomization

A final set of analyses was performed to investigate the effect on inter-rater and internal consistencies of dichotomizing the items into sufficient/not-sufficient (1,0) ratings. These are presented in Table 5. Dichotomization seemed to have little effect on inter-rater consistency, except among the correlations for the two and four item tests. These tended to be lowered somewhat, with a few increasing instead. The alphas also stayed similar for the larger tests, while both increasing and decreasing for the smaller tests Overall, these results suggest that the dichotomized ratings and the original four point ratings are equivalently reliable.

TABLE 4

ADJUSTED COEFFICIENT ALPHAS BASED ON INDICATORS, ESTIMATED FOR EQUAL TEST LENGTHS OF 10 ITEMS

		COEFFICIENT ALPHA				ADJUSTED COEFFICIENT ALPHA		
TEST	Ni	R1	R2	RS	NI	R1	R2	RS
T1 (All items)	20	.85	.88	.87	10	.74	.79	.77
T2A (Topic 1)	10	.89	.79	.88	10	.89	.79	.88
T2B (Topic 2)	10	.72	.79	.77	10	.72	.79	.77
T3A (Task 1)	10	.90	.92	.92	10	.90	.92	.92
T3B (Task 3)	10	.59	.83	.77	10	.59	.83	.77
T4A (Tpc 1, Tsk 1)	5	.95	.91	.95	10	.97	.95	.97
T4B (Tpc 1, Tsk 3)	5	.89	.72	.84	10	.94	.84	.91
T4C (Tpc 2, Tsk 1)	5	.86	.81	.87	10	.92	.90	.93
T4D (Tpc 2, Tsk 3)	5	.87	.83	.87	10	.93	.91	.93
T5A (Indicator 1)	4	.25	.69	.59	10	.45	.85	.78
T5B (Indicator 2)	4	.38	.51	.41	10	.61	.72	.63
T5C (Indicator 3)	4	.25	.20	.19	10	.45	.38	.37
T5D (Indicator 5)	4	.55	.64	.66	10	.75	.82	.83
T5E (Indicator 6)	4	.31	.77	.54	10	.53	.89	.75
T6A (Tsk 1, Ind 1)	2	.17	.88	.86	10	.51	.97	.97
T6B (Tsk 1, Ind 2)	2	.45	.54	.42	10	.80	.85	.78
T6C (Tsk 1, Ind 3)	2	.32	.86	.62	10	.70	.97	.89
T6D (Tsk 1, Ind 5)	2	.75	.70	.80	10	.94	.92	.95
T6E (Tsk 1, Ind 6)	2	.64	.71	.66	10	.90	.92	.91
T6F (Tsk 3, Ind 1)	2	.64	.73	.36	10	.90	.93	.74
T6G (Tsk 3, Ind 2)	2	.12	.62	.38	10	.41	.89	.75
T6H (Tsk 3, Ind 3)	2	.48	.28	0	10	.82	.66	0
T6I (Tsk 3, Ind 5)	2	.16	.60	.34	10	.49	.88	.72
T6J (Tsk 3, Ind 6)	2	0	.71	0	10	0	.92	0

TABLE 5

MEAN ITEM P-VALUES, COEFFICIENT ALPHAS, AND INTER-RATER CORRELATIONS FOR TESTS BASED ON DICHOTOMIZED INDICATORS

TEST	Ni	MEAN ITEM P VALUES R1	R2	RS	COEFFICIENT ALPHA R1	R2	RS	INTR RATR CORR
T1 (All items)	20	.23	.20	.22	.86	.88	.88	.53
T2A (Topic 1)	10	.27	.25	.26	.88	.83	.89	.56
T2B (Topic 2)	10	.20	.15	.18	.67	.68	.67	.53
T3A (Task 1)	10	.23	.13	.18	.91	.84	.90	.53
T3B (Task 3)	10	.24	.27	.26	.59	.82	.76	.57
T4A (Tpc 1, Tsk 1)	5	.26	.12	.19	.97	.96	.98	.56
T4B (Tpc 1, Tsk 3)	5	.28	.38	.33	.86	.70	.81	.38
T4C (Tpc 2, Tsk 1)	5	.20	.14	.17	.91	.13	.81	.61
T4D (Tpc 2, Tsk 3)	5	.20	.16	.18	.75	.82	.82	.82
T5A (Indicator 1)	4	.25	.22	.24	.51	.73	.67	.45
T5B (Indicator 2)	4	.25	.20	.23	.35	.32	.24	.34
T5C (Indicator 3)	4	.35	.32	.34	0	.51	.32	.48
T5D (Indicator 5)	4	.13	.10	.11	.81	.34	.77	.98
T5E (Indicator 6)	4	.20	.15	.18	.32	.75	.49	.40
T6A (Tsk 1, Ind 1)	2	.20	.15	.18	.55	.78	.93	.42
T6B (Tsk 1, Ind 2)	2	.30	.15	.23	.09	0	0	.39
T6C (Tsk 1, Ind 3)	2	.30	.20	.25	.09	.64	.53	.36
T6D (Tsk 1, Ind 5)	2	.20	.10	.15	.64	0	.46	.83
T6E (Tsk 1, Ind 6)	2	.15	.05	.10	.78	?	.44	.36
T6F (Tsk 3, Ind 1)	2	.30	.30	.30	.09	.69	.55	.64
T6G (Tsk 3, Ind 2)	2	.20	.25	.23	0	.53	.09	.30
T6H (Tsk 3, Ind 3)	2	.40	.45	.43	0	.16	.00	.67
T6I (Tsk 3, Ind 5)	2	.05	.10	.08	?	?	?	.67
T6J (Tsk 3, Ind 6)	2	.25	.25	.25	0	.53	0	.15

APPENDIX B:
SSI-EM SCORING MATERIALS

APPENDIX B:

SSI-EM SCORING MATERIALS

Candidate: ____ Scored by: ____ Score: _____

Scoring Form
Lesson Planning: Fractions

I. Components of the Lesson

a) The lesson on simplifying fractions
_____ 1. student activity
_____ 2. more than one representation of the content
_____ 3. candidate's mathematical accuracy
_____ 4. development of major idea: simplifying fractions

b) The 3 lesson sequence
_____ 5. emphasis on factoring
_____ 6. flow of the three lesson sequence
_____ 7. amount of practice in simplifying fractions
(Also consider the teacher's response to the question about the homework that s/he would assigned.)

c) Beginning of the lesson
_____ 8. clear introduction to simplifying fractions
_____ 9. ability to motivate students

d) Important features
_____ 10. factoring or greatest common factor
_____ 11. what simplifying means
_____ 12. one other idea related to simplifying fractions

e) Difficult features
_____ 13. knowing when the simplification is complete
_____ 14. finding factors or greatest common factor

II. Section Three: STUDENTS

a) Prior student knowledge

_____ 15. division, general concept of fractions,
_____ 16. factoring, and/or equivalent fractions
_____ 17. (3 out of this list of 4)

III. VIGNETTES

a) 4/20 = 1/5

_____ 18. appropriateness of response to student
_____ 19. use of alternative representation(s)

b) 8/18 divided by 4/4

_____ 20. appropriateness of response to student

Candidate ____________ Scorer _____________ Score ____________

Topic Sequencing: Fractions
SCORING FORM

Based on the candidate's responses to the questions in parts A through D rate the following categories:

RECORD THE CANDIDATE'S SORT:

* Candidate is able to accurately define:
_________ 1. FRS
_________ 2. LCM
_________ 3. TFA & TFO
_________ 4. CM
_________ 5. the other 12 concepts

* Candidate accurately perceives the significance of _______ to the overall topic of fractions:
_________ 6. LCM

_________ 7. CM

_________ 8. The candidate makes appropriate analogies (or accurate and understandable explanations) for the fraction concepts.

_________ 9. The candidate provides an appropriate explanation for why common denominators are needed for addition and subtraction.

_________ 10. The candidate addresses conceptual understandings in the card sort, especially when discussing multiplication of fractions.

_________ 11. Candidate, at some point in the interview, gives specific attention to the concept of fractions.

_________ 12. If the candidate chooses delete or add a topic, s/he provides either a pedagogical or mathematical justification.

Based on the candidate's responses to the questions in part E rate the following questions:

RECORD THE CANDIDATE'S SORT:

_________ 13. Candidate provides a good rationale for their distinction of difficult concepts.

_________ 14. SUBTRACTION OF MIXED NUMBERS WITH REGROUPING (SMR) is among the top two most difficult topics.

Candidate ____________ Scorer ______________ Score ____________

Topic Sequencing: Ratios
SCORING FORM

Based on the candidate's responses to the questions in parts A through D rate the following categories:

RECORD THE CANDIDATE'S SORT:

* Candidate is able to accurately define:

__________ 1. FRACTIONS AS A REGION/SET (FRS)
__________ 2. RATIO (RA)
__________ 3. PROPORTION (PR)
__________ 4. the other 14 concepts

* Candidate accurately perceives the significance of _______ to the overall topic of ratios:

__________ 5. COMPARISON OF NUMBERS (CN)
__________ 6. SCALE DRAWINGS (SD)
__________ 7. FINDING THE PERCENT OF A NUMBER (PN)

__________ 8. Candidate understands the relationship between proportions and equal ratios.

__________ 9. The candidate makes appropriate analogies (or accurate and understandable explanations) for the ratio concepts.

__________ 10. Candidate, at some point in the card sort, gives specific attention to the concept of ratios.

__________ 11. If candidate chooses to delete or add a topic, s/he provides either a pedagogical or mathematical justification.

Based on the candidate's responses to the questions in part E rate the following questions:

RECORD THE CANDIDATE'S SORT:

_________ 12. Candidate provides a good rationale for their distinction of difficult concepts.

_________ 13. CONVERTING FRACTIONS TO DECIMALS (FD) is among the top four most difficult topics.

_________ 14. Candidate demonstrates that conversion of decimals to fractions (DF) involves the determination of whether or not to reduce the fraction.

Candidate ______________ Scorer ___________ Score________________

Scoring Form
Shortcuts

'Gozinta' Method

__________ A. Identifies the shortcut's strengths. (only for a yes answer to #1)

__________ B. Identifies the shortcut's limitations.

__________ C. The candidate provides a suitable justification for teaching or not teaching the shortcut.

__________ D. Candidate describes appropriate ways to facilitate proper use of this method.

__________ E. Candidate describes good alternatives/complements for teaching the same idea that is incorporated in the shortcut.

__________ F. The candidate provides a mathematical rationale for why the shortcut does or does not work.

__________ G. The candidate properly identifies whether or not the shortcut always works.

__________ H. The candidate properly identifies the mathematical concepts that are embedded in the shortcut.

__________ AVERAGE SCORE FOR THE 'Gozinta' Method

Scoring Form: Shortcuts
(Page 2)

'1-2-3' Shortcut

__________ A. Identifies the shortcut's strengths. (only for a yes answer to #1)

__________ B. Identifies the shortcut's limitations.

__________ C. The candidate provides a suitable justification for teaching or not teaching the shortcut.

__________ D. Candidate describes appropriate ways to facilitate proper use of this method.

__________ E. Candidate describes good alternatives/complements for teaching the same idea that is incorporated in the shortcut.

__________ F. The candidate provides a mathematical rationale for why the shortcut does or does not work.

__________ G. The candidate properly identifies whether or not the shortcut always works.

__________ H. The candidate properly identifies the mathematical concepts that are embedded in the shortcut.

__________ AVERAGE SCORE FOR THE 'Gozinta' Method

__________ AVERAGE SCORE FOR THE '1-2-3' Method

__________ Final Score for the shortcuts exercise

Candidate ___________ Scorer ______________ Score __________

SCORING FORM
Lesson Planning: Ratios

I. Components of the lesson

a) The lesson percents and fractions

__________ 1. student activity
__________ 2. candidate's mathematical accuracy
__________ 3. development of major idea(s): %s and fractions
__________ 4. candidate's discussion of content attends to both procedures and conception

b) The 3 lesson sequence

__________ 5. emphasis on the mechanisms of conversion
__________ 6. flow of the three lesson sequence
__________ 7. amount of student practice (Also consider the teacher's response to the question about the homework that s/he would assign.)

c) Beginning of the lesson

__________ 8. clear introduction
__________ 9. ability to motivate students

d) Important/difficult features

__________ 10. simplification of fractions to lowest terms
__________ 11. percent signifies an amount out of 100
__________ 12. The same number can be expressed as both a fraction and a percent. Percents can be equal to fractions - even though they look different.

II. Section Three: STUDENTS

a) Prior student knowledge

__________ 13. division, fractions, percents,
__________ 14. factoring, equivalent fractions,
__________ 15. and/or simplest form
__________ 16. (4 out of this list of 6)

III. VIGNETTES

a) 2/1 = 200%, student thinks that 100% is largest percent

__________ 17. appropriateness for students

b) student converts 2/50 to 2%

__________ 18. appropriateness for students

c) student can't begin to convert 2/50 into a fraction

__________ 19. appropriateness for students

APPENDIX C:

ELEMENTARY EDUCATION EXAM CONSTRUCT VALIDITY

APPENDIX C:

ELEMENTARY EDUCATION EXAM CONSTRUCT VALIDITY

The data available for analysis were summary statistics from the Spring, 1989 California pilot test. These consisted of mean item scores (mean p-values) within content area subtests for 462 of the 480 teachers. From these means were calculated descriptive statistics (means, standard deviations, and minimum and maximum values) for the full group of teachers, as well as for breakdowns by undergraduate major, student status, ethnicity, and gender. Correlations within the full group also were provided among the six subtests, and between the subtests and grade point average, teaching status, SAT-Verbal, and SAT-Math.

These data and analyses cannot be used to address important technical issues such as test reliability, item or subset bias, or content validity. They do offer, however, a starting point for consideration of the construct validity of the test as a whole and of the six subject area subtests.

Correlations

The correlations among the subtests ranged from .09 (not significant) for Math with Other Subjects, to .42 (p<.0001) for Math with Pedagogy. The median between-subtest correlation was .27. All subtests, except Other Subjects, correlated significantly (p<.01) with GPA. No test correlated significantly with current teacher status (yes/no), including Pedagogy. SAT-Verbal correlated significantly with Social Studies and pedagogy, but not with any of the other subtests including Language Arts. SAT-Math correlated most highly with Pedagogy, and also correlated significantly with Math and Science.

The pattern of correlations suggest that overall competence level makes a relatively large contribution to the teachers' performances across all of the subtests. For example, Math correlates well, as would be expected, with SAT-M and with Science, however, it correlates even better with Language Arts, Pedagogy, and Social Studies. Further, Pedagogy is not significantly correlated with teacher status; instead, it correlates most highly with Math, LA, and SAT-M. This contraindicates the use of separate subtest scores, and weakens the interpretation of either subtest or full test scores as measures of pedagogical capabilities.

Mean Comparisons

The Other Subjects and Science subtests were the easiest, with mean p-values across teachers of .74 and .71. The most difficult was Math, with a mean p-value of .66. Standard deviations ranged between .13 and .19, so that the difference between the easiest and hardest subtest means was approximately one half of a standard deviation.

The breakdowns by undergraduate major show little interaction between major and subtest content area in determining mean performance; i.e. the relative order of performance level is roughly the same across all of the five specific subtest areas, regardless of the match between major and subtest. The items were most difficult for Education majors who score lowest on all subtests, including pedagogy. Liberal arts and English majors also score relatively low across all subtests. Conversely, business, science, and social studies majors always score above the mean, with science majors highest on Math and Science subtests, social science majors highest on Language Arts, and business major highest on Pedagogy and Social Studies. These results further support the correlational evidence against the construct validity of pedagogical interpretations, or use of separate subtests.

Analysis of Pedagogy means by student status found sophomores, juniors, and seniors averaging below the mean, and fifth year students, graduate students, and those not enrolled at or above the mean; with juniors lowest and those not enrolled highest.

Analyses of the item means by ethnicity reveal whites scoring on average from half to one standard deviation above blacks for all subtests. Asians and Hispanics tend to fall between the two, except that Asians score lowest of all groups on Pedagogy and Social Science and score as high as the white group on science. Gender differences are small, except for males scoring noticeably above females on Science and Social Studies and slight below on Pedagogy. These differences represent group mean differences in test performance within the sample tested. They do not address the question of test bias.

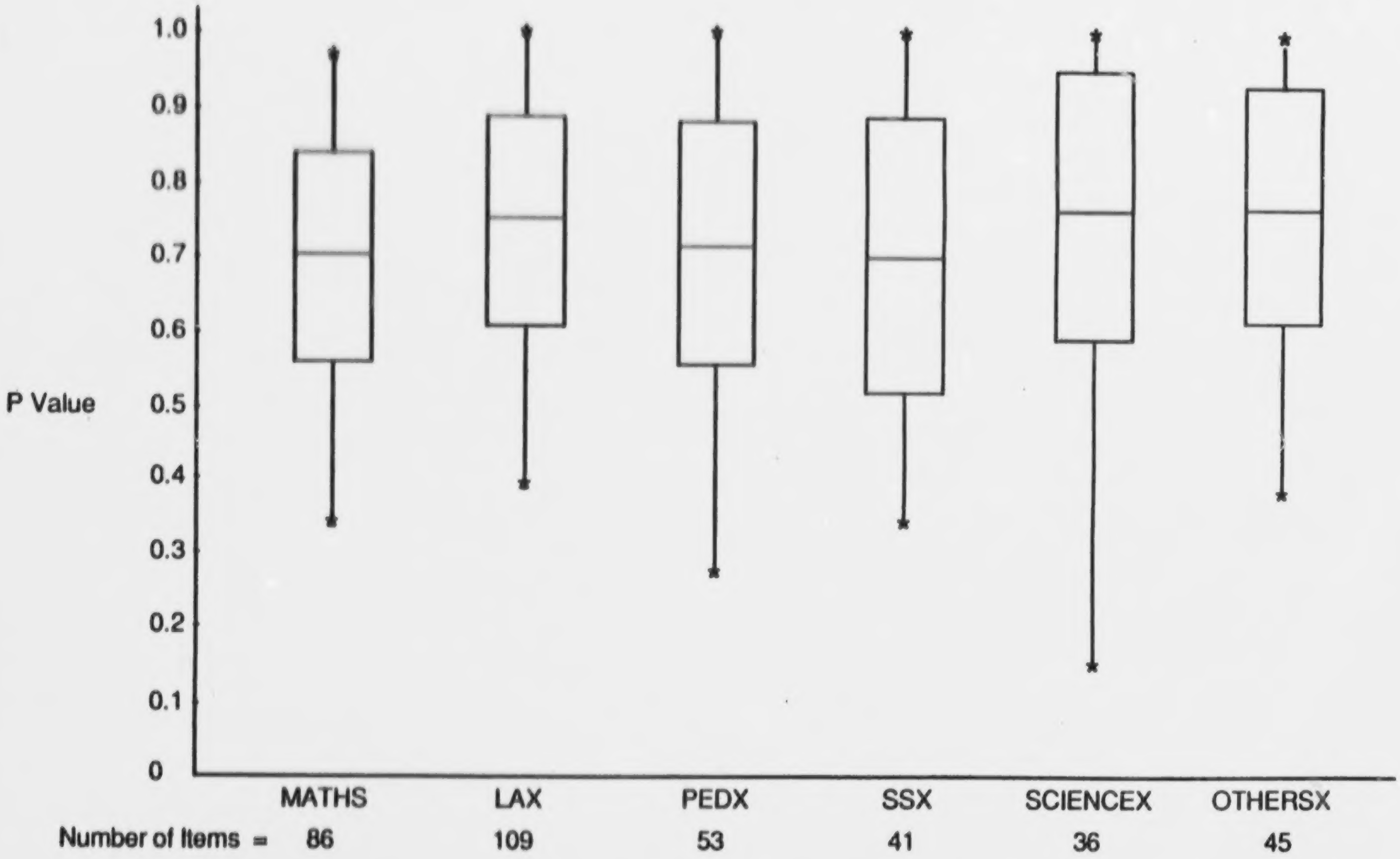
Further Analyses

Further analyses to address test reliability and item bias will require the individual teacher-item data. Once these data are obtained, coefficient alphas should be calculated to evaluate the internal consistency of the test and subtests.

FIGURE 1

ELEMENTARY EDUCATION EXAM

Sample: California Pilot Test Analysis Sample, Social Sciences Majors (N=45)

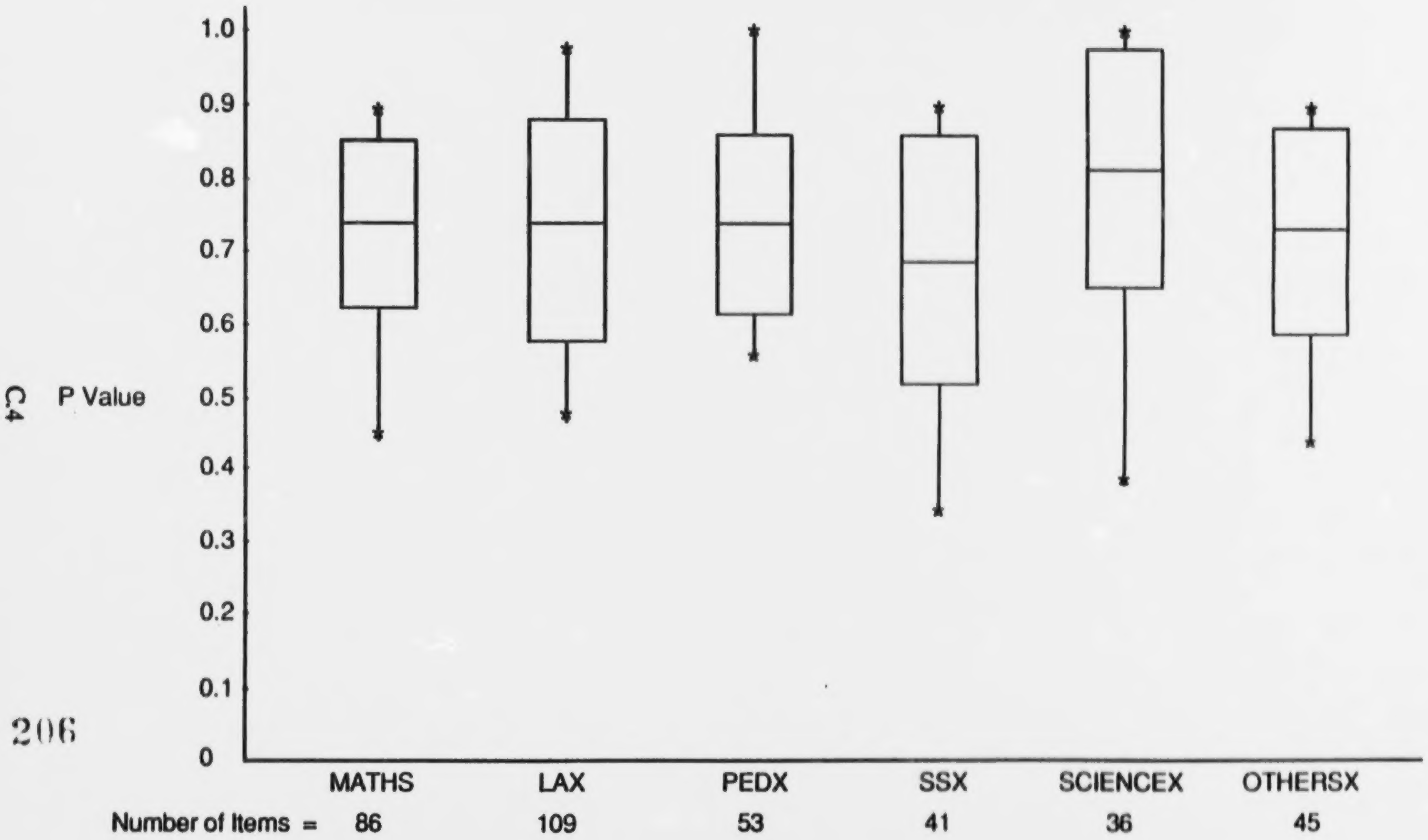


Bars represent means (over teachers) plus and minus one standard deviation of item means (over subject areas); asterisks represent minimum and maximum teachers' item means (over subject areas).

FIGURE 2

ELEMENTARY EDUCATION EXAM

Sample: California Pilot Test Analysis Sample, Science Majors (N=13)

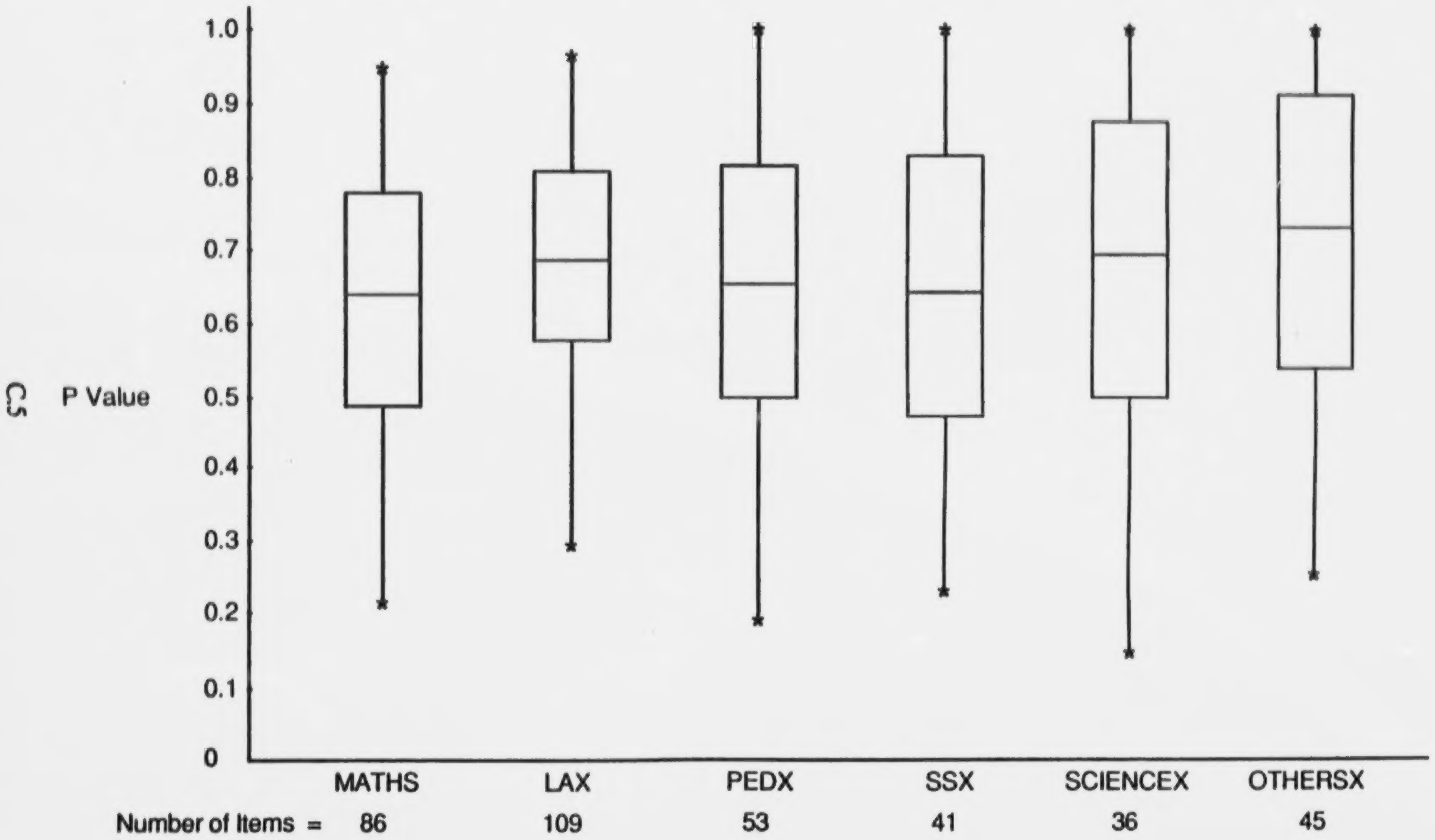


Bars represent means (over teachers) plus and minus one standard deviation of item means (over subject areas); asterisks represent minimum and maximum teachers' item means (over subject areas).

FIGURE 3

ELEMENTARY EDUCATION EXAM

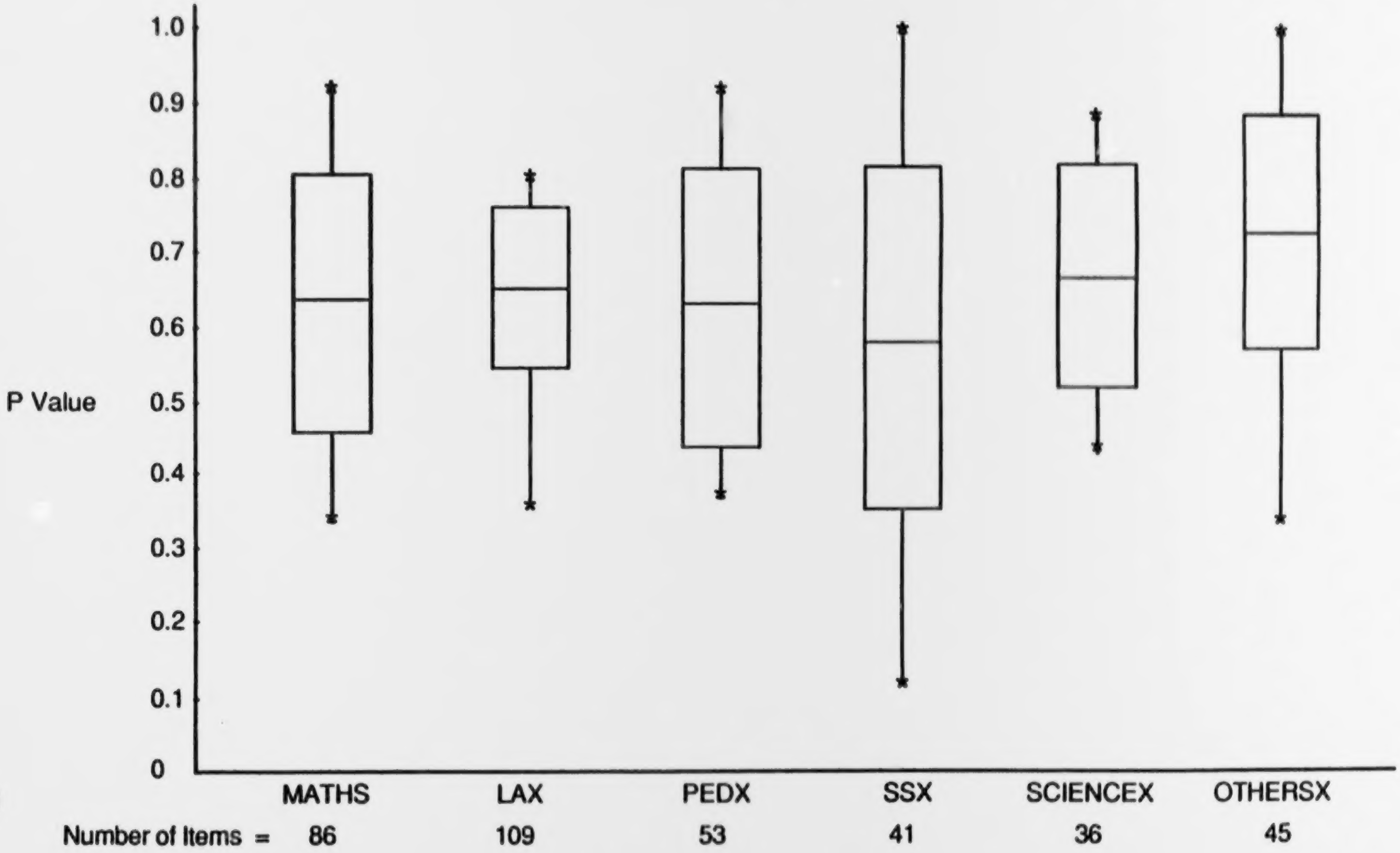
Sample: California Pilot Test Analysis Sample, Liberal Arts Majors (N=283)



Bars represent means (over teachers) plus and minus one standard deviation of item means (over subject areas); asterisks represent minimum and maximum teachers' item means (over subject areas).

FIGURE 4

ELEMENTARY EDUCATION EXAM
Sample: California Pilot Test Analysis Sample, Education Majors (N=20)



Bars represent means (over teachers) plus and minus one standard deviation of item means (over subject areas); asterisks represent minimum and maximum teachers' item means (over subject areas).